ارزیابی تأثیر رویکردهای مختلف انتخاب ویژگی بر مدل‌سازی جریان رودخانه با استفاده از الگوریتم‌های یادگیری ماشین

مریم میربیک سبزواری[۱]، حسن ترابی پوده[۲*]، محمدباقر دولتشاهی[۳]، امیرحمزه حقی‌آبی[۴]و عباس ملکی[۵]

۱- دانشجوی دکتری سازه‌های آبی، گروه مهندسی آب، دانشکده کشاورزی، دانشگاه لرستان، ایران.

۲*- نویسنده مسئول، استاد گروه مهندسی آب، دانشکده کشاورزی، دانشگاه لرستان، ایران، torabi.ha@lu.ac.ir

۳- دانشیار گروه مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه لرستان، ایران.

۴- استاد گروه مهندسی آب، دانشکده کشاورزی، دانشگاه لرستان، ایران.

۵- دانشیار گروه مهندسی آب، دانشکده کشاورزی، دانشگاه لرستان، ایران.

دریافت: بازنگری: پذیرش: انتشار:

چکیده

پیش‌بینی دقیق جریان رودخانه در سامانه‌های هیدرولوژیکی، به‌دلیل پیچیدگی روابط غیرخطی و تغییرات زمانی جریان، همچنان یک چالش اساسی محسوب می‌شود. الگوریتم‌های یادگیری ماشین به‌دلیل توانایی در مدل‌سازی روابط غیرخطی، به‌طور گسترده برای این منظور به‌کار می‌روند. با این حال، تعداد زیاد متغیرهای ورودی می‌تواند موجب افزایش پیچیدگی مدل و کاهش دقت پیش‌بینی شود. از این‌رو، انتخاب ویژگی، نقش مهمی در بهبود عملکرد مدل‌ها دارد. هدف این پژوهش بررسی تأثیر راهبردهای مختلف انتخاب ویژگی بر عملکرد مدل‌های یادگیری ماشین در پیش‌بینی ماهانه جریان رودخانه کشکان-پلدختر است. بدین منظور، دو رویکرد انتخاب ویژگی شامل روش مبتنی بر اهمیت ویژگی‌ها (SFM) و حذف بازگشتی ویژگی مبتنی بر اعتبارسنجی متقاطع (RFECV) همراه با سازوکار انتخاب ویژگی درونی برخی الگوریتم‌ها به‌کار گرفته شد و عملکرد الگوریتم‌های جنگل تصادفی (RF)، تقویت گرادیان شدید (XGBoost)، رگرسیون لاسو (LASSO) و رگرسیون بردار پشتیبان (SVR) برای پیش‌بینی دبی جریان بررسی و در مراحل آموزش و آزمون با شاخص‌های R^2، RMSE، MAE و KGE ارزیابی و با مدل پایه کلیماتولوژی مقایسه گردید. نتایج نشان داد روش انتخاب ویژگی تأثیر قابل توجهی بر عملکرد مدل‌ها دارد. مدل RF-SFM بهترین عملکرد را در فاز آزمون ارائه داد (R^2=0.926 و KGE=0.934) که بیانگر توانایی SFM در شناسایی متغیرهای مؤثر و بهبود دقت پیش‌بینی است. در مقابل، مدل RF-RFECVبا حذف تدریجی متغیرهای کم‌اهمیت، ساختاری ساده‌تر و پایدارتر ارائه داد که برای تحلیل‌های علمی و تفسیرپذیر مناسب‌تر است. همچنین XGBoost مستعد بیش‌برازش بوده و SVR ضعیف‌ترین عملکرد را در این مجموعه داده نشان داد. مقایسه نتایج با مدل پایه کلیماتولوژی نیز برتری الگوریتم‌های یادگیری ماشین در بازنمایی الگوهای پیچیده هیدرولوژیکی را نشان داد. در مجموع، انتخاب آگاهانه روش انتخاب ویژگی نقش مهمی در تعادل بین دقت پیش‌بینی و سادگی مدل داشته و می‌تواند در بهبود پیش‌بینی‌های عملیاتی جریان و پشتیبانی تصمیم‌گیری در مدیریت منابع آب بسیار مؤثر باشد.

کلید واژه‌ها: پیش‌بینی جریان، انتخاب ویژگی، جنگل تصادفی، یادگیری ماشین، مدل پایه کلیماتولوژی.

مقدمه

پیش‌بینی دقیق جریان رودخانه یکی از مؤلفه‌های اساسی در مدیریت منابع آب و برنامه‌ریزی سامانه‌های هیدرولوژیکی به‌شمار می‌رود. اطلاع از رفتار زمانی و مکانی جریان رودخانه، نقش کلیدی در بهره‌برداری بهینه از مخازن سدها، مدیریت خشکسالی، کنترل سیلاب، تأمین آب شرب، کشاورزی و تولید انرژی برق‌آبی ایفا می‌کند. با توجه به افزایش تقاضای آب و اثرات تغییرات اقلیمی، عدم قطعیت در الگوهای هیدرولوژیکی افزایش یافته و نیاز به روش‌های پیش‌بینی قابل اعتماد بیش از پیش احساس می‌شود. از این رو، توسعه مدل‌هایی که قادر به شبیه‌سازی و پیش‌بینی دقیق جریان رودخانه و و بهبود تصمیم‌گیری‌های مدیریتی باشند به‌عنوان یکی از چالش‌های مهم در مطالعه‌های مهندسی آب و هیدرولوژی مطرح است. مدل‌های فیزیکی و آماری سنتی که به‌طور گسترده برای پیش‌بینی جریان رودخانه مورد استفاده قرار گرفته‌اند، با وجود ارائه نتایج قابل قبول در بسیاری از مطالعه‌ها، با محدودیت‌های متعددی مواجه هستند. مدل‌های فیزیکی معمولاً نیازمند شناخت دقیق فرآیندهای حاکم بر حوضه آبریز و دسترسی به داده‌های گسترده و با کیفیت بالا هستند که در بسیاری از مناطق در دسترس نیست. از سوی دیگر، مدل‌های آماری کلاسیک مانند رگرسیون خطی و مدل‌های سری زمانی، عمدتاً بر فرضیات خطی بودن، ایستایی داده‌ها و توزیع نرمال متکی بوده و توانایی محدودی در مدل‌سازی روابط غیرخطی و پیچیده میان متغیرهای هیدرولوژیکی

دارند. این محدودیت‌ها به‌ویژه در شرایط حدی نظیر سیلاب‌ها و دوره‌های خشکسالی منجر به کاهش دقت پیش‌بینی شده و ضرورت به‌کارگیری رویکردهای پیشرفته‌تر را آشکار ساخته است. در سال‌های اخیر، روش‌های داده‌محور و یادگیری ماشین به‌دلیل توانایی بالا در استخراج الگوهای غیرخطی و پیچیده، به‌طور گسترده در پیش‌بینی دبی جریان مورد توجه قرار گرفته‌اند (Mosavi et al.,2018; Shen, 2018). مطالعه‌های اولیه در این حوزه عمدتاً بر مدل‌های آماری سنتی و رگرسیونی متمرکز بودند. با این حال، ناتوانی این روش‌ها در بازنمایی دینامیک‌های پیچیده و رخدادهای حدی، کارایی آن‌ها را به‌طور قابل توجهی محدود ساخت (Marco et al., 1993; Kisi, 2015). الگوریتم‌های یادگیری ماشین بدون نیاز به فرضیات سخت‌گیرانه درباره ساختار داده‌ها، قادر به یادگیری روابط پنهان بین متغیرهای ورودی و خروجی بوده و انعطاف‌پذیری بالایی در مواجهه با رفتارهای غیرایستا از خود نشان می‌دهند. شبکه‌های عصبی مصنوعی (Artificial Neural Networks: ANN) به‌عنوان یکی از نخستین ابزارهای یادگیری ماشین برای پیش‌بینی دبی و سیلاب به‌کار گرفته شدند. نتایج پژوهش‌های متعدد نشان داده است که این شبکه‌ها قادر به مدل‌سازی روابط پیچیده بین بارش، متغیرهای هیدرولوژیکی و دبی خروجی با دقت قابل قبول هستند. با این حال، مسائلی نظیر حساسیت به ساختار شبکه، نیاز به حجم بالای داده‌های آموزشی و خطر بیش‌برازش از محدودیت‌های اساسی این روش به شمار می‌روند (Maier et al., 2010; (Yaseen et al., 2015).

در ادامه توسعه روش‌های داده‌محور، الگوریتم‌های مبتنی بر یادگیری با حاشیه بیشینه نظیر ماشین بردار پشتیبان (Support Vector Regression: SVR) به‌طور گسترده در پیش‌بینی جریان رودخانه مورد توجه قرار گرفتند. مطالعه‌های متعدد نشان داده‌اند که SVR به‌دلیل برخورداری از اصل کمینه‌سازی ریسک ساختاری، در مجموعه داده‌های نسبتاً کوچک و پرنویز می‌تواند از پایداری بالاتری نسبت به شبکه‌های عصبی مصنوعی برخوردار باشد (Kisi et al., 2015). با این حال، عملکرد این روش به‌شدت به انتخاب تابع هسته، پارامترهای تنظیمی و مقیاس‌بندی داده‌ها وابسته است و عدم تنظیم مناسب این عوامل می‌تواند به ناپایداری و کاهش قابلیت تعمیم منجر شود (Zounemat-Kermani et al., 2021). به‌عنوان نمونه Yu et al. (2006) در پیش‌بینی کوتاه‌مدت سطح جریان رودخانه نشان دادند که SVM در افق‌های چندساعته تا چندروزه نسبت به ANN دقت بالاتری دارد. Liu et al. (2014) نیز برتری SVM را نسبت به مدل‌های خطی کلاسیک در پیش‌بینی بلندمدت جریان گزارش کردند. با وجود این دستاوردها، محدودیت SVR در بازنمایی رفتارهای حدی و پیچیدگی تنظیم ابرپارامترها، کاربرد عملی آن را در بسیاری از حوضه‌های پیچیده محدود ساخته است. با گسترش رویکردهای تجمیعی (Ensemble) مدل‌های مبتنی بر درخت تصمیم به‌ویژه جنگل تصادفی (Random Forest) به یکی از ابزارهای اصلی پیش‌بینی هیدرولوژیکی تبدیل شدند. ساختار تجمیعی RF مبتنی بر نمونه‌گیری بوت‌استرپ و انتخاب تصادفی ویژگی‌ها، این الگوریتم را در برابر بیش‌برازش مقاوم ساخته و امکان مدل‌سازی روابط غیرخطی و برهم‌کنش‌های پیچیده میان متغیرهای اقلیمی و هیدرولوژیکی را فراهم می‌کند. نتایج مطالعه‌های متعدد نشان داده است که RF در مقایسه با ANN و SVR از دقت و پایداری بالاتری در پیش‌بینی دبی جریان برخوردار است (Breiman, 2001; Mosavi et al.,2018; Tyralis et al., 2019). به موازات آن، الگوریتم‌های تقویت گرادیان مانند XGBoost و LightGBM به‌عنوان نسل پیشرفته‌تری از مدل‌های درختی معرفی شده‌اند که با بهره‌گیری از بهینه‌سازی گرادیانی، توانایی بالایی در استخراج الگوهای پیچیده داده‌ها دارند (Chen and Guestrin, 2016). پژوهش‌های اخیر نشان می‌دهند که XGBoost به‌ویژه در مسائل پیش‌بینی دبی جریان رودخانه عملکرد بسیار رقابتی و با دقت بالایی از خود نشان می‌دهد. با این وجود، حساسیت این مدل‌ها به تنظیم ابرپارامترها و خطر بیش‌برازش در مجموعه داده‌های محدود هیدرولوژیکی از چالش‌های اصلی آن‌ها به شمار می‌رود (Fang et al., 2024). علاوه بر توسعه مدل‌های منفرد، پژوهش‌های متعددی بر ترکیب روش‌های یادگیری ماشین با رویکردهای هیبریدی تمرکز کرده‌اند. Tongal و Booij (2018) نشان دادند که جداسازی جریان به مؤلفه‌های پایه و سطحی و استفاده هم‌زمان از متغیرهای هواشناسی می‌تواند عملکرد SVR، ANN و RF را به‌طور محسوسی بهبود بخشد. Fathian et al. (2019) نیز با ترکیب مدل‌های غیرخطی نظیر RF و MARS با ساختارهای سری زمانی، به دقت بالاتری نسبت به مدل‌های مستقل دست یافتند و ترکیب RF-SETAR را به‌عنوان کاراترین گزینه معرفی کردند. در رویکردی مشابه Ni et al.(2020) با تلفیق خوشه‌بندی GMM و XGBoost نشان دادند که تفکیک رژیم‌های هیدرولوژیکی پیش از مدل‌سازی می‌تواند قابلیت پیش‌بینی جریان ماهانه را افزایش دهد. Szczepanek(2022) در مقایسه مدل‌های تقویت گرادیان برای پیش‌بینی جریان روزانه، بر نقش تعیین‌کننده کیفیت داده آموزشی و تنظیم ابرپارامترها در عملکرد نهایی تأکید کرد. مطالعه‌های اخیر نشان داده‌اند که کیفیت انتخاب متغیرهای ورودی نقشی تعیین‌کننده در عملکرد مدل‌های یادگیری ماشین برای پیش‌بینی جریان رودخانه دارد. Fang et al. (2024) با ارزیابی پنج راهبرد انتخاب متغیر مبتنی بر جنگل تصادفی (RF-IVS) در پیش‌بینی جریان یک ماهه و در چارچوب مدل‌های RF، SVR، فرآیند گاوسی (GP) و LSTMنشان دادند که روش‌های مبتنی بر عملکرد پیش‌بینی، نسبت به رویکردهای کلاسیک مانند همبستگی جزیی پیرسون و اطلاعات متقابل شرطی از دقت بالاتری برخوردارند. نتایج این پژوهش بیانگر آن بود که ترکیب RF با راهبرد انتخاب رو به جلو و مدل GP قادر است تعادلی مناسب میان کاهش ابعاد فضای ویژگی و حفظ اطلاعات هیدرولوژیکی برقرار سازد و به بهبود معنی‌دار دقت

پیش‌بینی منجر شود. در مطالعه‌های منطقه‌ای نیز کارایی متفاوت الگوریتم‌های یادگیری ماشین گزارش شده است. برای مثال Elhallaoui Oueldkaddour (2025) در پژوهشی بر روی داده‌های جریان ماهانه در اقلیم نیمه خشک مراکش با مقایسه سه مدل SVR، RF و XGBoost نشان داند که SVR در آن مجموعه داده خاص عملکرد بهتری ارائه کرده و می‌تواند در شرایط کمبود داده جایگزین مناسبی برای مدل‌های پیچیده‌تر باشد. این نتایج بار دیگر تأکید می‌کند که برتری یک الگوریتم به ویژگی‌های حوضه، مقیاس زمانی و کیفیت متغیرهای ورودی وابسته است و نمی‌توان یک مدل واحد را برای تمامی شرایط بهینه دانست. فراتر از انتخاب الگوریتم، پژوهش‌های نظری نیز بر اهمیت انتخاب ویژگی (Feature Selection) در بهبود عملکرد و تعمیم‌پذیری مدل‌های یادگیری ماشین تأکید کرده‌اند. Guyon و Elisseeff (2003) نشان دادند که استفاده از تعداد زیاد متغیرهای هم‌بسته لزوما موجب بهبود مدل نمی‌شود و حتی می‌تواند به کاهش قابلیت تعمیم منجر گردد. در مطالعات اخیر حوزه هیدرولوژی نیز بر ضرورت به‌کارگیری راهبردهای هدفمند انتخاب ویژگی برای افزایش تفسیرپذیری و قابلیت تعمیم مدل‌های داده‌محور تأکید شده است. روش‌های انتخاب ویژگی توکار مانند SelectFromModel (SFM) و روش‌های پوششی مانند حذف بازگشتی مبتنی بر اعتبارسنجی متقاطع (Recursive Feature Elimination using Cross-Validation: RFECV) به‌دلیل بهره‌گیری از ساختار درونی مدل، امکان شناسایی مجموعه‌ای حداقلی و در عین حال مؤثر از متغیرها را فراهم می‌کنند. کاربرد این رویکردها در مطالعه‌های هیدرولوژیکی به‌طور فزاینده‌ای گسترش یافته است. گزارش‌ها حاکی از آن است که اگرچه ممکن است کاهش اندکی در دقت ایجاد شود اما در مقابل پایداری ساختاری و قابلیت تعمیم مدل به‌طور قابل توجهی بهبود می‌یابد؛ ویژگی‌هایی که برای کاربردهای عملی نظیر پیش‌بینی دقیق جریان رودخانه و مدیریت منابع آب از اهمیت بالایی برخوردار هستند (Tyralis et al., 2019; Nearing et al., 2021). در حوزه ترکیب انتخاب ویژگی با مدل‌های عمیق Ma et al. (2024) با تلفیق RFECV با LSTM نشان دادند که حذف تکراری متغیرهای غیرمؤثر می‌تواند به مجموعه‌ای بهینه از ویژگی‌ها منجر شده و دقت پیش‌بینی رواناب ماهانه را به‌طور چشمگیری افزایش دهد. همچنین Zhang et al. (2026) در قالب مدل SFM-MF برای حوضه‌های کوچک و دارای داده محدود گزارش کردند که استفاده از ویژگی‌های برگزیده توسط SFM عملکرد مدل‌های پایه نظیر RF و XGBoost را به شکل محسوسی ارتقا می‌دهد. این یافته‌ها بیانگر آن است که انتخاب ویژگی نه صرفاً یک مرحله پیش‌پردازش، بلکه بخشی جدایی‌ناپذیر از فرآیند مدل‌سازی هیدرولوژیکی است. با وجود پیشرفت‌های قابل توجه، مرور ادبیات نشان می‌دهد که هنوز توافق جامعی درباره برتری رویکردهای مختلف انتخاب ویژگی در پیش‌بینی دبی جریان وجود ندارد و عملکرد نهایی مدل‌ها به تعامل پیچیده‌ای میان نوع الگوریتم، روش انتخاب ویژگی، راهبرد اعتبارسنجی و ویژگی‌های اقلیمی–هیدرولوژیکی حوضه وابسته است. به‌ویژه مقایسه دو رویکرد پرکاربرد SFM و RFECV در مقیاس دبی ماهانه و در حوضه‌های کوهستانی ایران کمتر مورد توجه قرار گرفته است؛ در حالی‌که این مقیاس به‌دلیل نوسانات شدید و وجود تأخیرهای زمانی، چالش‌های ویژه‌ای را برای مدل‌های داده‌محور ایجاد می‌کند. با وجود این پیشرفت‌ها، یک خلأ پژوهشی جدی همچنان پابرجاست: اغلب مطالعه‌ها یا بر افزایش دقت متمرکز بوده‌اند و یا بر کاهش ابعاد، اما کمتر پژوهشی به‌طور هم‌زمان به مبادله بنیادین میان دقت، پایداری و تفسیرپذیری پرداخته است؛ به‌ویژه در مقیاس دبی ماهانه حوضه‌های کوهستانی که رفتار غیرایستا، تأخیرهای زمانی و رخدادهای حدی، چالشی فراتر از داده‌های ماهانه ایجاد می‌کند. در ایران نیز علی‌رغم اهمیت عملی پیش‌بینی سیلاب و مدیریت مخازن، مقایسه SFM و RFECV در چارچوب مدل‌های مختلف یادگیری ماشین تقریباً مغفول مانده است.

در پژوهش حاضر با هدف پر کردن این شکاف، عملکرد چهار الگوریتم یادگیری ماشین شامل RF، XGBoost، LASSO و SVR برای پیش‌بینی دبی ماهانه ایستگاه کشکان–پلدختر ارزیابی شده و دو راهبرد انتخاب ویژگی SFM و RFECV در چارچوب اعتبارسنجی متقاطع k بخشی همراه با رویکردهای درونی انتخاب ویژگی مورد مقایسه قرار گرفته‌اند. نوآوری اصلی این مطالعه نه صرفاً معرفی یک مدل جدید، بلکه بازتعریف معیار برتری مدل‌ها بر اساس تعادل میان دقت پیش‌بینی و قابلیت تعمیم عملیاتی است. همچنین به‌منظور سنجش ارزش افزوده مدل‌های یادگیری ماشین، نتایج با یک مدل پایه (Baseline Model) کلیماتولوژی مقایسه شده است که مدل پایه کلیماتولوژی امکان سنجش ارزش افزوده واقعی یادگیری ماشین را فراهم می‌سازد. هدف اصلی این مطالعه، شناسایی ساختاری است که ضمن ارائه دقت پیش‌بینی بالا، از پایداری و قابلیت تعمیم مناسبی برخوردار بوده و بتواند به‌عنوان ابزاری قابل اعتماد در مدیریت عملیاتی منابع آب به کار گرفته شود.

مواد و روش‌ها

منطقه مورد مطالعه

حوضه آبریز رودخانه کشکان با وسعت ۹۵۰۲/۷ کیلومترمربع در ناحیه جنوب غربی ایران واقع شده است. این حوضه از زیرحوضه‌های درجه سه حوضه آبریز کرخه بوده و حدود ۱۲/۵ درصد از مساحت آن را شامل می‌شود. از نظر موقعیت جغرافیایی این حوضه در محدوده '۱۲ °۴۷ تا '۵۹ °۴۸ طول شرقی و '۸ °۳۳ تا '۲ °۳۴ عرض شمالی در بخش میانی کوه‌های زاگرس قرار دارد. این حوضه از لحاظ تقسیمات سیاسی در استان لرستان قرار گرفته است و ۳۳ درصد از مساحت کل استان را به خود اختصاص داده است (Lorestan Regional Water Company, 2014). رودخانه کشکان به‌عنوان یکی از مهم‌ترین شاخه‌های تأمین‌کننده جریان ورودی به رودخانه کرخه، نقش کلیدی در مدیریت منابع

آب منطقه، کنترل سیلاب و تأمین نیازهای کشاورزی و زیست محیطی ایفا می‌کند. ایستگاه هیدرومتری کشکان-پلدختر به‌عنوان ایستگاه مبنای این پژوهش، خروجی اصلی حوضه را پایش می‌کند و به‌دلیل قرارگیری در پایین‌دست حوضه، نماینده مناسبی از رفتار یکپارچه هیدرولوژیکی آن محسوب می‌شود. موقعیت محدوده مطالعاتی در شکل (۱) نشان داده شده است.

داده‌ها

داده‌های مورد استفاده در این مطالعه شامل داده‌های دو ایستگاه هیدرومتری و نه ایستگاه هواشناسی می‌باشند که موقعیت مکانی این ایستگاه‌ها در شکل (۱) و مشخصات آن‌ها در جداول (۱) و (۲) ارائه شده است. برای مدل‌سازی پارامترهای بارش از متغیرهای P(t)، P(t-1)، P(t-2) و P(t-3) استفاده شد که به‌ترتیب بیان‌گر میانگین بارش ماه جاری، یک، دو و سه ماه قبل در ایستگاه هیدرومتری بالادست کاکارضا است که به‌عنوان ورودی در نظر گرفته شدند. متغیر هدف مدل، دبی ماهانه در ایستگاه هیدرومتری کشکان-پلدختر در بازه زمانی دی ماه سال ۱۳۴۸ تا شهریور ماه ۱۴۰۳ استفاده گردید. انتخاب تأخیرهای زمانی بر مبنای ملاحظات هیدرولوژیکی و تحلیل وابستگی زمانی جریان انجام شد. استفاده از دبی ایستگاه کاکارضا به‌عنوان ورودی، با هدف انتقال اطلاعات هیدرولوژیکی بخش بالادست به مدل صورت گرفت؛ زیرا این متغیر نمایانگر پاسخ تجمعی فرآیندهای بارش-رواناب در زیرحوضه‌های بالادست بوده و می‌تواند الگوهای مکانی-زمانی جریان را در مقیاس حوضه به‌طور مؤثرتری منعکس نماید. به‌منظور جلوگیری از ورود عدم قطعیت ناشی از داده‌های ناقص، تمامی رکوردهایی که در هر یک از متغیرهای ورودی یا خروجی دارای داده مفقود بودند از مجموعه داده حذف گردیدند. پس از اعمال این کنترل کیفی، تعداد ۶۳۳ نمونه هم‌زمان برای فرآیند آموزش و آزمون مدل‌ها در دسترس قرار گرفت که از این داده‌ها برای آموزش و تست مدل‌ها استفاده شد. کلیه مراحل پیش‌پردازش داده‌ها، پیاده‌سازی مدل‌ها، انتخاب ویژگی و ارزیابی عملکرد با استفاده از زبان برنامه‌نویسی پایتون (Python) انجام شد. در این مطالعه پنج ساختار مدل‌سازی شامل RF-RFECV، RF-SFM، SVR-SFM، LASSO و XGBoost به‌منظور ارزیابی جامع توانایی روش‌های یادگیری ماشین در پیش‌بینی دبی جریان توسعه یافتند. برای مدل RF دو رویکرد مستقل انتخاب ویژگی شامل RFECV و SFM به‌کار گرفته شد تا تأثیر روش انتخاب ویژگی بر دقت و پایداری مدل بررسی شود. در مدل SVR انتخاب ویژگی صرفاً از طریق روش SFMانجام گرفت. رگرسیون LASSO به‌دلیل برخورداری از سازوکار ذاتی انتخاب ویژگی از طریق صفر شدن ضرایب متغیرهای کم اهمیت، بدون روش انتخاب ویژگی خارجی اجرا شد.

مدل XGBoost نیز به‌صورت مستقیم و بدون اعمال روش انتخاب ویژگی خارجی مورد استفاده قرار گرفت، زیرا این الگوریتم به‌واسطه ساختار درختی و مکانیزم کاهش گرادیان، به‌طور ضمنی قابلیت ارزیابی اهمیت متغیرها را داراست. به‌منظور ارزیابی میزان بهبود حاصل از استفاده از روش‌های یادگیری ماشین، عملکرد تمامی مدل‌ها با یک مدل پایه مقایسه شد تا ارزش افزوده هر ساختار مدل‌سازی به‌صورت کمی بررسی شود.

مدل جنگل تصادفی همراه با حذف بازگشتی ویژگی‌ها (RF-RFECV)

الگوریتم جنگل تصادفی یک روش تجمیعی مبتنی بر درخت تصمیم است که پیش‌بینی نهایی را از طریق میانگین‌گیری بر روی مجموعه‌ای از درخت‌های تصمیم انجام می‌دهد. این الگوریتم با استفاده از نمونه‌گیری بوت‌استرپ (bootstrap) و انتخاب تصادفی ویژگی‌ها در هر گره، واریانس مدل را کاهش داده و در برابر بیش‌برازش مقاوم است و توانایی بالایی در مدل‌سازی روابط غیرخطی و پیچیده هیدرولوژیکی دارد (Breiman, 2001; Hastie et al., 2009). در مدل RF-RFECV به‌منظور کاهش ابعاد فضای ویژگی و انتخاب مؤثرترین متغیرهای ورودی از رویکرد انتخاب ویژگی بازگشتی مبتنی بر اعتبارسنجی متقاطع (RFECV) استفاده شد (Pedregosa et al., 2011). ایده اصلی روش حذف بازگشتی ویژگی‌ها نخستین بار توسط Guyon et al. (2002) مطرح شد که در آن ویژگی‌ها به‌صورت تدریجی و بر اساس میزان اهمیت حذف می‌شوند تا زیرمجموعه‌ای بهینه از متغیرها حاصل گردد. در چارچوب RFECV، عملکرد مدل در هر مرحله با استفاده از اعتبارسنجی متقاطع k بخشی ارزیابی می‌شود تا انتخاب ویژگی بر مبنای قابلیت تعمیم مدل صورت گیرد و از انتخاب بیش برازش شده جلوگیری شود (Kuhn and (Johnson, 2013). در این فرآیند ابتدا مدل RF با تمامی متغیرها آموزش داده می‌شود و بر اساس معیار اهمیت ویژگی‌ها، کم‌اثرترین متغیر حذف می‌گردد (Breiman, 2001). این روند به‌صورت بازگشتی ادامه می‌یابد تا ترکیبی از ویژگی‌ها که بیشترین کارایی پیش‌بینی را در اعتبارسنجی متقاطع ارائه می‌دهد انتخاب شود (Pedregosa et al., 2011). استفاده از ارزیابی متقاطع در چرخه حذف ویژگی، خطر نشت اطلاعات (information leakage) میان داده‌های آموزش و آزمون را کاهش داده و منجر به بهبود قابلیت تعمیم مدل به داده‌های مستقل می‌شود (Kuhn and Johnson, 2013). بنابراین انتخاب مجموعه‌ای حداقلی و بهینه از متغیرهای ورودی، منجر به افزایش پایداری، کاهش پیچیدگی مدل و بهبود قابلیت تعمیم آن در پیش‌بینی دبی جریان می‌شود.

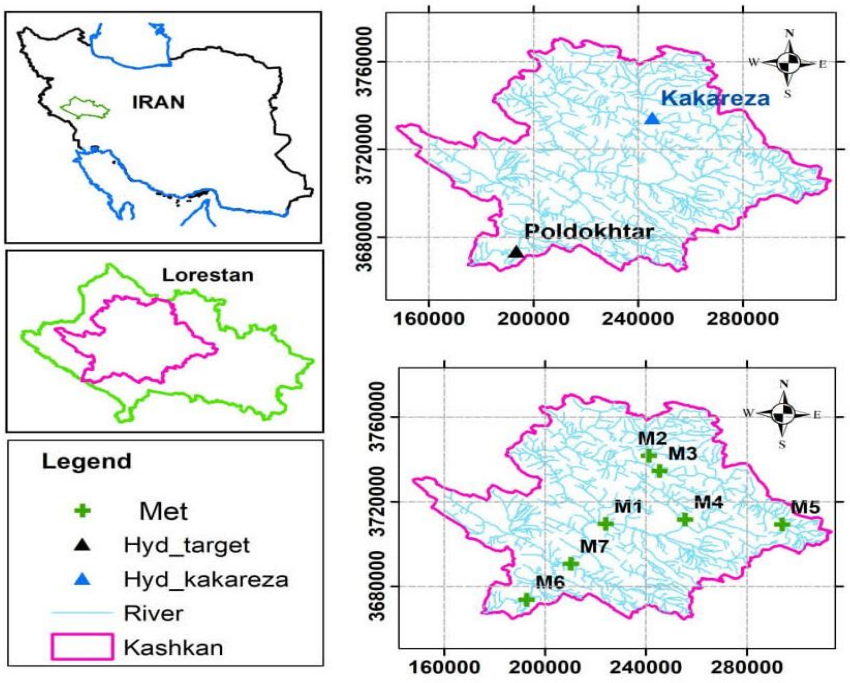


Fig. 1- Study area

شکل ۱- محدوده مورد مطالعه

جدول ۱- مشخصات ایستگاه‌های هواشناسی

Table 1- Characteristics of meteorological stations

Row	Station	Elevation (m)	Longitude (°E)	Latitude (°N)	Maximum (mm)	Mean (mm)	Standard Deviation (mm)	Map Label
1	Veysian	1008	46°59′09″	33°31′18″	355	23.2	53.30	M1
2	Sarabseidali	1570	47°09′59″	33°49′44″	334	24	52.94	M2
3	Kakareza	1531	47°12′54″	33°46′13″	425.4	28	61.70	M3
4	Darreh Tang	1711	47°14′30″	33°57′37″	458.5	24.5	53.23	M4
5	Khorramabad	1291	47°19′29″	33°32′29″	280	22	47.78	M5
6	Noorabad	1788	47°43′18″	33°31′16″	357	24	49.71	M6
7	Kuhdasht	1197	48°12′56″	33°33′24″	298	16.4	48.40	M7
8	Poldokhtar	668	48°19′03″	33°13′46″	333	15	45.46	M8
9	Afrine	820	48°29′24″	33°23′32″	411.5	19	53.16	M9

جدول ۲- مشخصات ایستگاه‌های هیدرومتری

Table 2- Characteristics of hydrometric stations

Station	River	Area (km^2)	Elevation (m)	Longitude (°E)	Latitude (°N)	Minimum (m^3/s)	Maximum (m^3/s)	Mean (m^3/s)	Standard Deviation (m^3/s)
Kakareza	Harod	1154	1527	47°12′54″	33°46′12″	0.158	213.48	4.72	15.83
Poldokhtar	Kashkan	9268	670	48°19′03″	33°12′36″	0	849.81	28.87	59.99

در مدل RF-RFECV، الگوریتم جنگل تصادفی به‌عنوان مدل پیش‌بینی و هم به‌عنوان هسته ارزیابی اهمیت ویژگی‌ها مورد استفاده قرار گرفت (Tyralis et al., 2019). ابر پارامترهای مدل شامل تعداد درخت‌ها، عمق بیشینه و حداقل نمونه در برگ‌ها با مقادیر پیش‌فرض یا تنظیم جزیی تنظیم شدند و معیار خطای

میانگین مربعات (MSE) به‌عنوان شاخص ارزیابی در مراحل حذف ویژگی به‌کار رفت (Hastie et al., 2009). خروجی این فرآیند، مجموعه‌ای فشرده از متغیرهای هیدرولوژیکی و اقلیمی بود که ضمن حفظ بیشترین اطلاعات، پیچیدگی مدل را کاهش داد و زمینه را برای دست‌یابی به مدلی پایدارتر و تفسیرپذیرتر فراهم ساخت.

مدل جنگل تصادفی همراه با انتخاب ویژگی توکار (RF-SFM)

در این مدل از روش انتخاب ویژگی توکار مبتنی بر اهمیت متغیرها (SFM) در ترکیب با الگوریتم جنگل تصادفی استفاده شد. در رویکردهای توکار، فرآیند انتخاب ویژگی به‌صورت درونی و هم‌زمان با آموزش مدل انجام می‌شود و برخلاف روش‌های پوششی، نیازمند اجرای مکرر مدل در چرخه‌های بازگشتی نیست (Guyon and Elisseeff, 2003). الگوریتم جنگل تصادفی به‌دلیل برخورداری از سازوکار برآورد اهمیت ویژگی‌ها، گزینه‌ای مناسب برای پیاده‌سازی SFM به‌شمار می‌آید (Breiman, 2001). در جنگل تصادفی، اهمیت هر متغیر معمولاً بر اساس میزان کاهش ناخالصی گره‌ها (Gini/Variance Reduction) یا بر مبنای روش جایگشت (Permutation Importance) محاسبه می‌شود. متغیرهایی که سهم بیشتری در کاهش خطای پیش‌بینی دارند، اهمیت بالاتری دریافت می‌کنند و می‌توانند به‌عنوان زیرمجموعه بهینه ویژگی‌ها انتخاب شوند (Breiman, 2001; Hastie et al., 2009; Liaw and Wiener, 2002). در چارچوب SFM، پس از آموزش اولیه مدل RF با تمامی متغیرها، ویژگی‌هایی که اهمیت آن‌ها از یک آستانه مشخص کمتر بود حذف شده و مدل نهایی تنها با متغیرهای منتخب بازآموزی می‌شود (Pedregosa et al., 2011). این رویکرد ضمن حفظ توان پیش‌بینی، پیچیدگی مدل را کاهش داده و به بهبود تفسیرپذیری کمک می‌کند (Molnar, 2020). مزیت اصلی مدل RF-SFM نسبت به روش‌های پوششی نظیر RFECV، کارایی محاسباتی بالاتر و پایداری بیشتر در مواجهه با داده‌های هم‌بسته و نویزی است، زیرا انتخاب ویژگی بر پایه ساختار درونی مدل و بدون نیاز به جستجوی تکراری در فضای ویژگی انجام می‌گیرد (Guyon and Elisseeff, 2003; Kuhn and Johnson, 2013). در این مدل، برای تعیین آستانه انتخاب ویژگی از معیار اهمیت مبتنی بر کاهش واریانس خطا استفاده شد و ابرپارامترهای مدل RF شامل تعداد درخت‌ها و عمق بیشینه با مقادیر پیش‌فرض یا تنظیم جزیی انتخاب شدند (Hastie et al., 2009). خروجی فرآیند RF-SFM، مجموعه‌ای بهینه از متغیرهای ورودی بود که بیشترین سهم را در پیش‌بینی دبی جریان داشت و مبنای توسعه مدل نهایی RF-SFM قرار گرفت.

مدل ماشین بردار پشتیبان همراه با انتخاب ویژگی مبتنی بر اهمیت متغیرها (SVR–SFM)

رگرسیون بردار پشتیبان (SVR) کاربرد ماشین بردار پشتیبان در مسائل رگرسیونی است که بر پایه اصل کمینه‌سازی ریسک ساختاری انجام می‌شود (Vapnik, 1995). با توجه به عملکرد پایدار و سهولت کار با SVR این روش به یکی از الگوریتم‌های کلاسیک یادگیری ماشین در پیش‌بینی جریان تبدیل شده است (Ikram et al. 2022 ؛Liu et al. 2014). تابع رگرسیونی SVR را می‌توان به‌صورت رابطه (۱) بیان کرد:

$$y=\omega\phi(x)+b \qquad (۱)$$

که ω و b به‌ترتیب وزن‌ها و بایاس هستند و $\phi(.)$ تابع نگاشت است که فضای ورودی x را به فضایی با بُعد بالاتر منتقل می‌کند. برای امکان‌پذیر کردن یک خطای از پیش تعیین شده ε در تابع رگرسیون، تابع هزینه ε-غیرحساس به‌صورت رابطه (۲) تعریف می‌شود.

$$L_\varepsilon(d,y) = |d-y|_\varepsilon = \max(0, |d-y| - \varepsilon) \qquad (۲)$$

تابع هزینه زمانی برابر با صفر است که اختلاف بین مقادیر هدف d و مقادیر خروجی y در محدوده خطای مجاز باشد. هدف SVR یافتن یک ابرصفحه بهینه است که تا حد امکان تخت (flat) بوده و مقدار تابع هزینه آن حداقل باشد (Smola and Schölkopf, 2004). از این رو، مسئله بهینه‌سازی می‌تواند به‌صورت رابطه (۳) نوشته شود:

$$\min f = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{N}|d-y|_\varepsilon \qquad (۳)$$

کمینه شدن عبارت جریمه $\|\omega\|^2$ تخت بودن تابع رگرسیون را تضمین می‌کند. N اندازه نمونه‌ها و C یک ثابت است که توازن بین تختی مدل و انحراف‌های بیش از آستانه از پیش تعیین شده را مشخص می‌کند (Hao and Bai 2023). از آنجا که SVR به‌طور ذاتی سازوکار مستقیمی برای معیار اهمیت ویژگی ندارد، در این مدل اهمیت ویژگی‌ها بر پایه یک مدل پشتیبان مبتنی بر درخت جنگل تصادفی استخراج و سپس در قالب چارچوب SFM برای انتخاب زیرمجموعه بهینه متغیرها به‌کار گرفته شد. در این فرآیند، ابتدا اهمیت نسبی تمامی متغیرها برآورد و ویژگی‌هایی که اهمیت آن‌ها کمتر از آستانه تعیین شده بود حذف گردید، سپس مدل SVR تنها با متغیرهای منتخب آموزش داده شد. چنین رویکردی در مطالعه‌های مختلف به‌عنوان یک راهکار متداول برای ترکیب مدل‌های غیرخطی فاقد معیار اهمیت درونی با روش‌های انتخاب ویژگی گزارش شده است (Guyon and Elisseeff, 2003). در نسخه‌های SVR با کرنل RBF، پارامترهای اصلی شامل C، ε و عرض کرنل (γ) هستند که کنترل گستردگی نگاشت داده‌ها به فضای با بُعد بالاتر را بر عهده دارند (Smola and Schölkopf, 2004). در این پژوهش انتخاب مناسب این پارامترها با مقادیر پیش‌فرض و آزمون و خطا انجام شد

تا عملکرد مدل بهینه گردد. به‌کارگیری SFM پیش از آموزش SVRبا کاهش ابعاد فضای ویژگی، حساسیت مدل به هم‌خطی متغیرها را کاهش داده و پایداری عددی آن را بهبود بخشید. اگرچه SVR در مجموعه داده‌های کوچک عملکرد قابل قبولی نشان می‌دهد، اما حساسیت بالای آن نسبت به انتخاب پارامترها و تابع هسته، پایداری آن را در مسائل پیچیده هیدرولوژیکی نسبت به مدل‌های تجمیعی کاهش می‌دهد.

رگرسیون لاسو (LASSO)

رگرسیون لاسو یک مدل خطی تعمیم‌یافته است که با اعمال جریمه بر ضرایب رگرسیون، برخی ضرایب را به صفر کاهش می‌دهد و بدین ترتیب امکان انتخاب ویژگی به‌صورت توکار فراهم می‌شود. در این روش ضرایب رگرسیونی β به‌گونه‌ای برآورد می‌شوند که مجموع مربعات باقیمانده کمینه گردد، مشروط بر آنکه مجموع مقادیر مطلق ضرایب از یک مقدار ثابت بیشتر نشود. برآورد ضرایب β معمولاً با روش حداقل مربعات محاسبه می‌شود که می‌تواند به‌صورت رابطه (۴) تعریف شود:

$$L(\beta)=\|Y-X\beta\| \qquad (۴)$$

که $L(\beta)$ تابع هزینه، Y بردار مقادیر هدف و X ماتریس متغیرهای ورودی‌ها می‌باشد. برای انتخاب پیش‌بینی‌کننده‌ها، یک عبارت جریمه (penalty term) به‌صورت رابطه (۵) به آن اضافه شده است:

$$L(\beta) = \|Y-X\beta\| + \lambda\beta \qquad (۵)$$

که در آن λ پارامتر جریمه مثبت است. این پارامتر توازن بین خطای برازش و ساده‌سازی مدل را کنترل می‌کند. به‌طوری‌که با افزایش مقدار λ میزان جریمه بر ضرایب بیشتر شده و مدل ساده‌تری با ضرایب کمتر حاصل می‌شود و تعداد بیشتری از ضرایب به سمت صفر میل می‌کنند (Kwon et al., 2017). زمانی‌که مدل لاسو، N جفت نمونه آموزشی را در نظر بگیرد مسئله برآورد ضرایب β می‌تواند به‌صورت رابطه (۶) فرموله شود:

$$\beta = \operatorname{argmin}\left\{\sum_{i=1}^{N}\left(y_i-\beta_0-\sum_{j=1}^{k}\beta_j x_{ij}\right)^2\right\} \quad s.t \quad \sum_{j=1}^{k}\beta_j \le \lambda \qquad (۶)$$

در تعیین مقدار پارامتر λ باید دقت کافی صورت گیرد، زیرا این پارامتر نقش تعیین‌کننده‌ای در حذف متغیرهای اضافی و حفظ پیش‌بینی‌کننده‌های مؤثر دارد و به‌طور هم‌زمان بر دقت برآورد ضرایب باقی‌مانده تأثیر می‌گذارد. در این روش، ضرایبی که به سمت صفر میل می‌کنند بیانگر حذف آن متغیرها از مدل نهایی بوده و بدین ترتیب مجموعه‌ای از پیش‌بینی‌کننده‌های مؤثر انتخاب می‌شود (Chu et al., 2018). رگرسیون لاسو به‌طور هم‌زمان عمل انتخاب ویژگی و برازش مدل را انجام می‌دهد و این ویژگی باعث افزایش تفسیرپذیری مدل و کنترل مؤثر بیش‌برازش می‌گردد.

الگوریتم تقویت گرادیان شدید (XGBoost)

XGBoost نسخه‌ای بهبودیافته از الگوریتم درخت تصمیم گرادیان تقویتی (Gradient-Boosting Decision Tree: GBDT) می‌باشد (Chen and Guestrin, 2016). یادگیری سریع و دقیق این روش باعث شده است که XGBoost به یک مدل برتر یادگیری ماشین در علوم داده از جمله مدل‌سازی هیدرولوژیکی-هواشناسی تبدیل شود (Ko ،Liu et al., 2022 et al., 2020 ،Potdar et al., 2021). الگوریتم در این الگوریتم، خروجی نهایی مدل، حاصل جمع پیش‌بینی‌های K درخت تصمیم است که هر کدام به‌صورت تدریجی به مدل اضافه می‌شوند و مقادیر پیش‌بینی شده آن را می‌توان به‌صورت رابطه (۷) نوشت:

$$y = \sum_{k=1}^{K} f_k(x) \qquad (۷)$$

که f_k متناظر با ساختار یک درخت با T برگ و وزن‌های برگ ω می‌باشد. یکی از تفاوت‌های اصلی XGBoost با GBDT این است که آن‌ها از توابع هدف متفاوتی استفاده می‌کنند. در مدل XGBoost یک عبارت جریمه نیز اضافه شده است که به‌صورت رابطه (۸) بیان می‌شود:

$$Obj = \sum_{i=1}^{N} l(y_i, d_i) + \sum_{k=1}^{K} \Omega(f_k) \qquad (۸)$$

که در آن l تابع هزینه بوده که اختلاف بین مقدار پیش‌بینی‌شده y و مقدار هدف d را اندازه‌گیری می‌کند. به‌منظور افزایش دقت، تابع هزینه با بسط تیلور مرتبه دوم تقریب زده می‌شود. عبارت دوم، بیانگر مجموع پیچیدگی هر تابع پایه f است که برای جلوگیری از بیش‌برازش به کار می‌رود و به‌صورت رابطه (۹) تعریف می‌شود:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2 \qquad (۹)$$

که γ و λ پارامترهای جریمه هستند و به‌ترتیب تعداد برگ‌ها T و بزرگی وزن‌های برگ ω را کنترل می‌کنند (Hao and Bai, 2023). الگوریتم XGBoost قابلیت شناسایی برهم‌کنش‌های پیچیده بین متغیرها را دارد. در این پژوهش، مدل XGBoost بدون استفاده از روش انتخاب ویژگی خارجی به کار رفت، زیرا این الگوریتم به‌طور ضمنی از طریق معیارهای اهمیت ویژگی و فرآیند تقویت گرادیان، نقش متغیرهای مؤثر را ارزیابی می‌کند. با وجود عملکرد رقابتی و پایداری بالای XGBoost در بسیاری از مطالعه‌های هیدرولوژیکی، حساسیت آن نسبت به تنظیم پارامترها و خطر بیش‌برازش از چالش‌های اصلی این الگوریتم محسوب می‌شود.

مدل کلیماتولوژی (Climatology)

در این پژوهش مدل کلیماتولوژی به‌عنوان یک مدل پایه برای ارزیابی عملکرد مدل‌ها در نظر گرفته شد. این مدل فرض می‌کند که مقدار پیش‌بینی شده دبی جریان برای هر بازه زمانی برابر با میانگین تاریخی جریان‌ها در همان روز یا ماه طی دوره آموزشی است. مدل کلیماتولوژی به‌دلیل سادگی و کاربرد گسترده به‌عنوان یک مرجع پایه در مطالعه‌های پیش‌بینی هیدرولوژیکی به‌کار می‌رود و امکان ارزیابی میزان بهبود عملکرد مدل‌های پیشرفته‌تر، از جمله روش‌های یادگیری ماشین را نسبت به یک پیش‌بینی مرجع فراهم می‌کند. در نتیجه، مقایسه نتایج مدل‌های توسعه یافته با این مدل پایه می‌تواند نشان‌دهنده میزان ارزش افزوده روش‌های پیشنهادی در پیش‌بینی جریان رودخانه باشد (Crochemore et al., 2017; Costa et al., 2021). در برخی مطالعه‌ها، روش‌های پایه دیگری مانند مدل تداوم (Persistence) یا مدل‌های آماری سری زمانی نیز برای مقایسه به‌کار گرفته می‌شوند که بررسی آن‌ها می‌تواند در مطالعه‌های آینده مدنظر قرار گیرد.

معیارهای ارزیابی

برای ارزیابی دقت و عملکرد مدل‌های پیشنهادی از شاخص‌های استاندارد شامل ضریب تعیین (Coefficient of Determination: R^2)، ریشه میانگین مربعات خطا (RMSE: Root Mean Squared Error)، میانگین قدر مطلق خطا (MAE: Mean Absolute Error) و ضریب کارایی کلینگ-گوپتا (KGE: Kling-Gupta Efficiency) استفاده گردید (Ghorbani et al. 2016؛ Gupta et al. 2009). بهترین مقادیر این معیارها به‌ترتیب یک، صفر، صفر، یک و یک هستند. به علاوه برای R^2، مقدار یک نشان‌دهنده دقت بالای پیش‌بینی و مقدار صفر نشان‌دهنده شکست مدل پیشنهادی در تعیین تغییرپذیری میانگین است (Nagelkerke, 1991). فرمول‌های مورد نیاز برای محاسبه شاخص‌های مورد نظر در معادله‌های (۱۰) تا (۱۳) ارائه شده‌اند. چارچوب کلی پژوهش نیز در شکل (۲) نشان شده است.

$$R^2 = \left[\frac{\sum_{i=1}^{n}(M_{oi}-\bar{M}_o)(M_{ei}-\bar{M}_e)}{\sqrt{\sum_{i=1}^{n}(M_{oi}-\bar{M}_o)^2 \sum_{i=1}^{n}(M_{ei}-\bar{M}_e)^2}}\right]^2 , 0 \le R^2 \le 1 \quad (۱۰)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(M_{ei} - M_{oi})^2} , 0 \le RMSE \le \infty \quad (۱۱)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|M_{ei} - M_{oi}| , \quad 0 \le MAE \le \infty \quad (۱۲)$$

$$KGE = 1 - \sqrt{(r-1)^2 + \left(\frac{\sigma_{ei}}{\sigma_{oi}} - 1\right)^2 + \left(\frac{\bar{M}_e}{\bar{M}_o} - 1\right)^2}$$
$$-\infty < KGE < 1 \quad (۱۳)$$

در روابط بالا، M_{oi} و M_{ei} به‌ترتیب مقادیر مشاهداتی و محاسباتی در گام زمانی iام، n تعداد گام‌های زمانی، $\bar{M}_o$ و $\bar{M}_e$ نیز به‌ترتیب میانگین مقادیر مشاهداتی و محاسباتی است. r ضریب همبستگی پیرسون بین مقادیر مشاهداتی و محاسباتی و $\frac{\sigma_{ei}}{\sigma_{oi}}$ نسبت انحراف معیار داده‌های محاسباتی به مشاهداتی است.

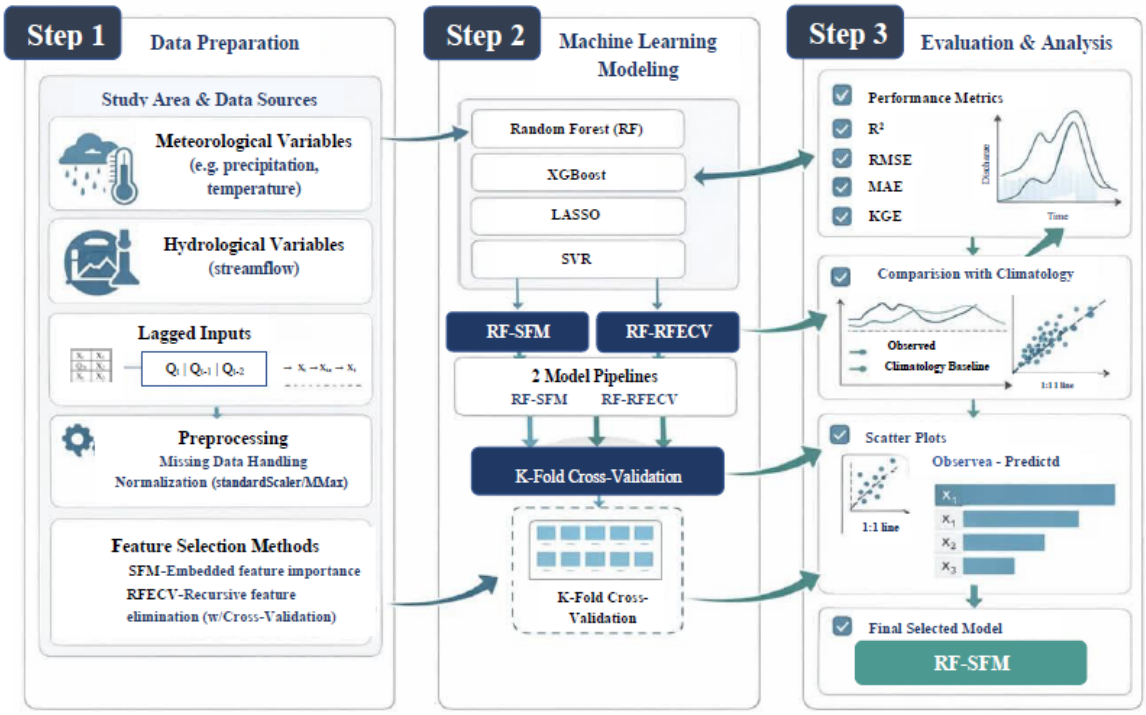


Fig. 2- Overall methodological framework for monthly streamflow prediction using machine learning models and feature selection approaches

شکل ۲- چارچوب کلی روش‌شناسی پیشنهادی برای پیش‌بینی دبی ماهانه با استفاده از مدل‌های یادگیری ماشین و روش‌های انتخاب ویژگی

نتایج و بحث

به‌منظور ارزیابی دقت، پایداری و قابلیت تعمیم مدل‌های مختلف در پیش‌بینی دبی جریان، عملکرد پنج ساختار یادگیری ماشین به‌همراه یک مدل پایه کلیماتولوژی مورد بررسی قرار گرفت. ارزیابی مدل‌ها در دو فاز آموزش و آزمون و با استفاده از معیارهای ارزیابی انجام شد و نتایج در جدول (۳) ارائه گردید. نتایج ارائه شده در جدول (۳) نشان می‌دهد که رفتار مدل‌ها تفاوت قابل توجهی در تعادل میان دقت و پایداری دارند. تمامی مدل‌های مبتنی بر یادگیری ماشین عملکردی به‌مراتب بهتر از مدل پایه کلیماتولوژی داشته‌اند. مقادیر منفی R^2 و KGE برای مدل کلیماتولوژی در فاز آزمون بیانگر توان بسیار محدود این مدل در بازتولید تغییرات دبی جریان بوده و نقش آن را صرفاً به‌عنوان یک خط مبنای مقایسه‌ای تأیید می‌کند که در ادامه به‌صورت تفصیلی مورد بحث قرار می‌گیرد.

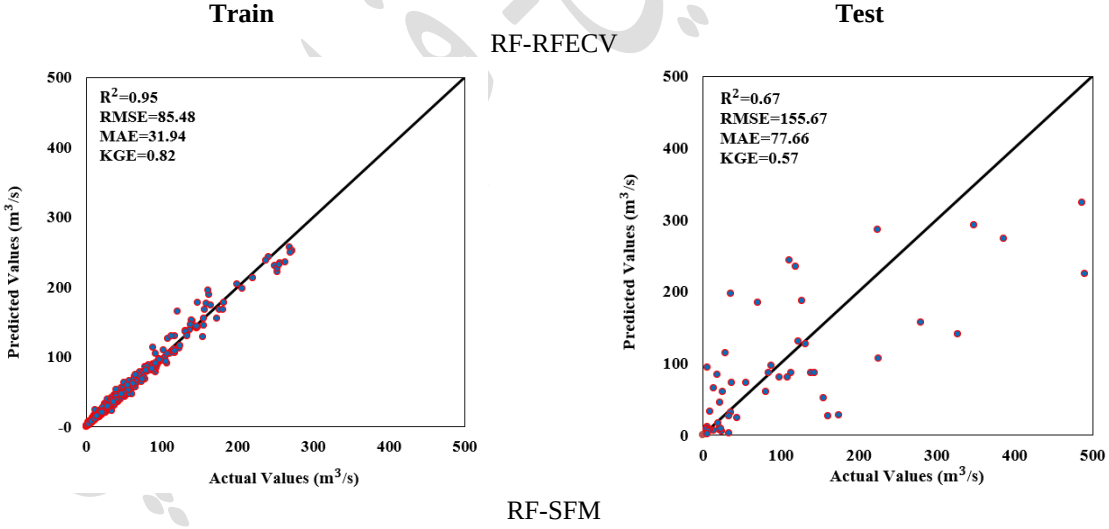
عملکرد مدل RF-RFECV

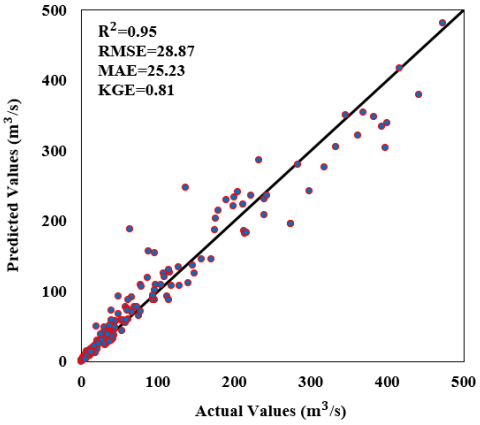
همان‌طور که در جدول (۳) نشان داده شده است مدل RF-RFECV در فاز آموزش از دقت بالایی برخوردار بوده به‌طوری‌که مقدار ضریب تعیین برابر با R^2=۰/۹۵۴ و مقدار خطای RMSE=۸۵/۴۷۶ و MAE=۳۱/۹۴۲ به‌دست آمده است. با این حال، در فاز آزمون افت محسوسی در عملکرد مدل مشاهده می‌شود و مقدار R^2به ۰/۶۷۵ کاهش یافته و RMSE و MAE به‌ترتیب به ۱۵۵/۶۷۳ و ۷۷/۶۶۵ افزایش یافته‌اند. همچنین کاهش شاخص KGE از ۰/۸۱۶ در فاز آموزش به ۰/۵۷۴ در فاز آزمون، بیانگر کاهش توان تعمیم مدل به داده‌های آزمون است. این اختلاف قابل توجه میان دو فاز نشان می‌دهد که انتخاب ویژگی مبتنی بر RFECV اگرچه به نقطه بهینه اعتبارسنجی متقاطع دست یافته است اما با حذف بخش قابل توجهی از متغیرهای ورودی، ساختاری بیش از حد ساده ایجاد کرده است. این نتیجه نشان می‌دهد که RFECV به‌طور آگاهانه به سمت مدل ساده‌تر و پایدارتر حرکت نموده و در ازای افزایش قابلیت تعمیم و کاهش وابستگی به متغیرهای همبسته، بخشی از دقت پیش‌بینی کاهش یافته است. اگرچه این رویکرد باعث کاهش پیچیدگی مدل و افزایش پایداری ساختاری می‌شود اما بخشی از اطلاعات کلیدی لازم برای بازنمایی دینامیک پیچیده جریان رودخانه را حذف کرده و در نتیجه دقت پیش‌بینی کاهش یافته است. افزایش حدود ۸۲ درصدی RMSE و بیش از دو برابر شدن MAE نسبت به فاز آموزش مؤید این مسئله است. همان‌طور که در شکل (۳) نشان داده شده است نمودار پراکنش عملکرد مدل RF-RFECV با نتایج عددی جدول (۳) همخوانی دارد. در فاز آموزش، نقاط به‌خوبی حول خط y=x متمرکز شده و شیب رگرسیون نزدیک به واحد است که با مقدار بالای R^2 مطابقت دارد. اما در فاز آزمون، پراکندگی نقاط افزایش یافته و فاصله از خط بهینه بیشتر می‌شود؛ به‌ویژه در مقادیر بالای دبی، مدل به برآورد کم تمایل نشان می‌دهد که با افزایش RMSE و کاهش

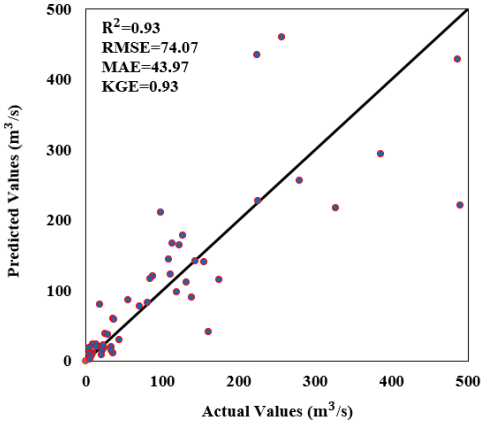
KGEدر جدول (۳) سازگار است. مدل RF-RFECV با زیرمجموعه‌ای محدود از ویژگی‌های منتخب و تنظیمات پیش‌فرض جنگل تصادفی آموزش دیده است در این مدل، تعداد درخت‌ها برابر ۱۰۰ (پیش‌فرض) در نظر گرفته شد. عمق بیشینه و حداقل نمونه در برگ با استفاده از آزمون و خطا برابر ۱۰ تعیین شد. مقدار k=5 برای اعتبارسنجی متقاطع نیز به‌کار رفت که هم تعادل مناسبی بین دقت و هزینه محاسباتی ایجاد می‌کند و هم در مطالعات مشابه رایج است. کاهش بعد فضای ویژگی از طریق RFECV اگرچه موجب کاهش هم‌خطی و افزایش پایداری عددی مدل شده اما ظرفیت مدل برای بازنمایی روابط غیرخطی میان بارش‌های دارای تأخیر زمانی و دبی ماهانه را محدود ساخته است. بنابراین، عملکرد این مدل، نمونه‌ای روشن از مبادله میان دقت و سادگی است؛ به‌گونه‌ای که رویکرد انتخاب ویژگی تهاجمی‌تر به کاهش محسوس توان پیش‌بینی در فاز آزمون منجر شده است.

عملکرد مدل RF-SFM

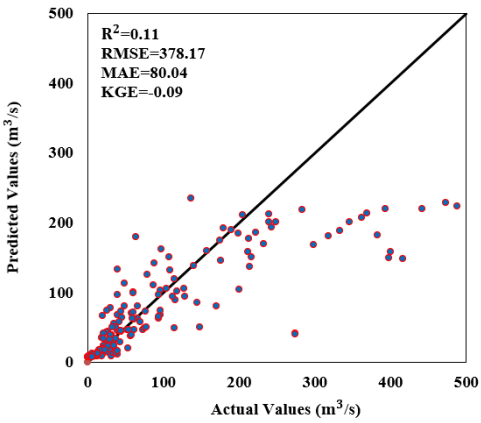
طبق جدول (۳) مدل RF-SFM بهترین عملکرد کلی را در میان تمامی مدل‌ها ارائه داده است. این مدل در فاز آزمون به مقادیر بسیار بالای R^2=۰/۹۲۶ و KGE=۰/۹۳۴ دست یافته است که نشان‌دهنده تطابق بسیار مناسب بین مقادیر پیش‌بینی شده و مشاهداتی از نظر همبستگی، واریانس و بایاس می‌باشد. مقادیر پایین خطا شامل RMSE=۷۴/۰۶۷ و MAE =۴۳/۹۶۷ بیانگر توان بالای مدل در بازتولید دامنه تغییرات دبی جریان و کاهش خطای پیش‌بینی است. فاصله کم میان معیارهای آموزش و آزمون نشان می‌دهد که مدل توانسته بدون بیش‌برازش، تعادلی مطلوب میان دقت و قابلیت تعمیم برقرار نماید. همان‌طور که در شکل (۳) نشان داده شده است پراکندگی نقاط در اطراف خط ۴۵ درجه در نمودار آزمون بیانگر توانایی مناسب مدل در شبیه‌سازی دبی جریان است. تمرکز نقاط حول خط y=x و شیب رگرسیون نزدیک به واحد، توانایی RF-SFM در بازنمایی رفتار واقعی جریان را نشان می‌دهد. مدل RF-SFM با استفاده از زیرمجموعه‌ای از ویژگی‌های منتخب و راهبرد توکار SFM آموزش دیده است. این سازوکار باعث شده که مدل ضمن بهره‌گیری از مجموعه‌ای غنی از متغیرهای ورودی، از پیچیدگی غیرضروری اجتناب کند و ظرفیت مناسبی برای بازنمایی روابط غیرخطی میان بارش‌های دارای تأخیر زمانی و دبی ماهانه داشته باشد. در این مدل، تعداد درخت‌ها برابر ۱۰۰ (پیش‌فرض)، عمق بیشینه و حداقل نمونه در برگ با استفاده از آزمون و خطا برابر ۱۰ و مقدار k نیز برابر پنج تعیین شد. بنابراین مدل RF-SFM نمونه‌ای موفق از ترکیب قدرت یادگیری مدل‌های تجمیعی با انتخاب ویژگی هدفمند است و می‌تواند به‌عنوان گزینه‌ای قابل اعتماد برای پیش‌بینی عملیاتی دبی جریان ماهانه پیشنهاد شود.

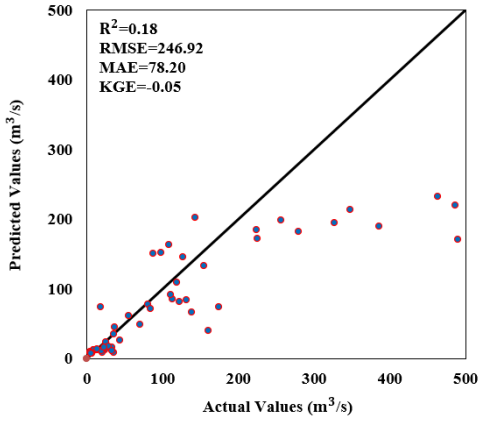




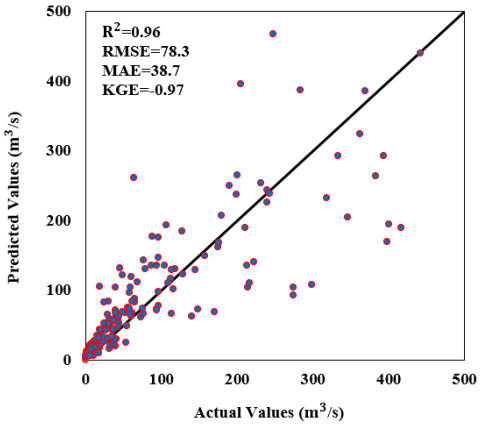


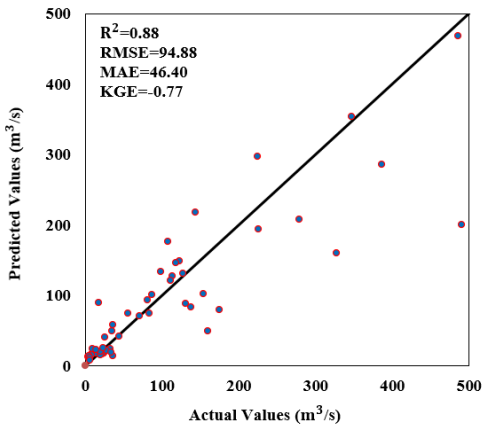
SVR-SFM





LASSO





XGBoost

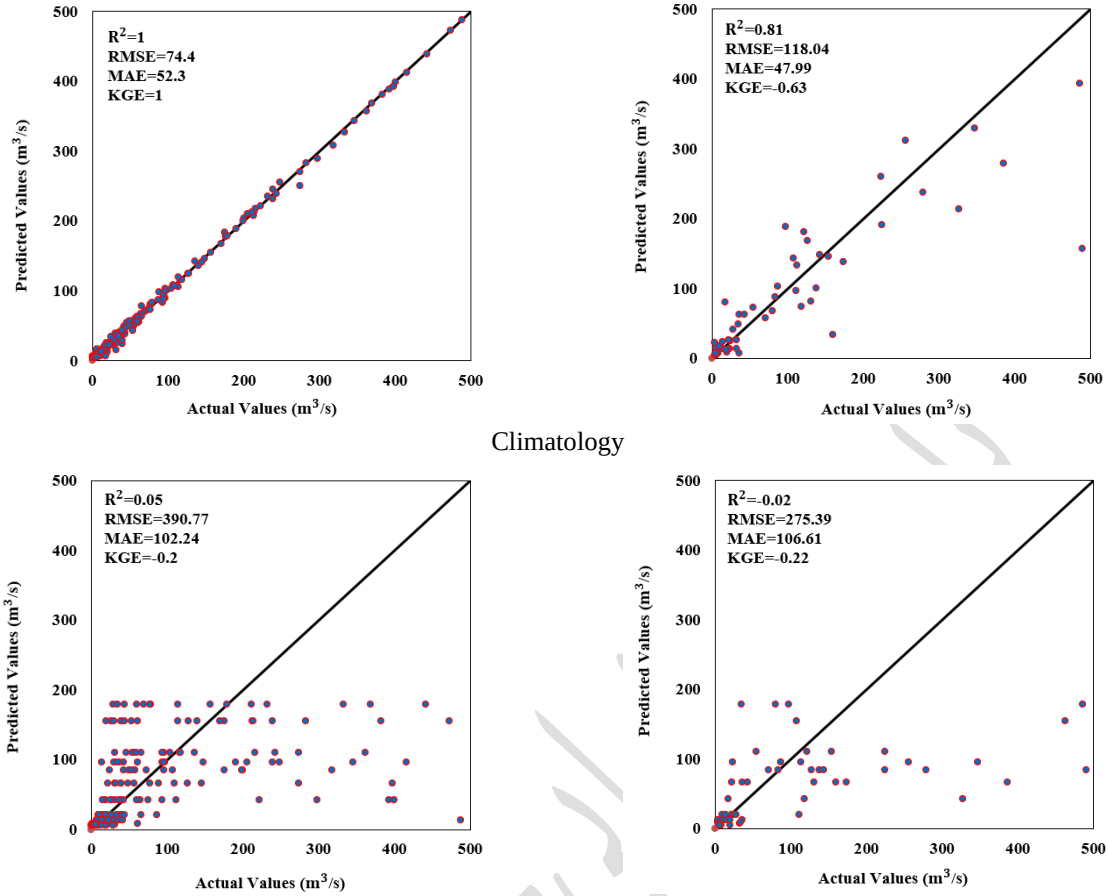


Fig. 3- Scatter plot of observed and simulated values from streamflow prediction models in the training and testing phases

شکل ۳- نمودار پراکنش مقادیر مشاهداتی و محاسباتی حاصل از مدل‌های پیش‌بینی دبی جریان در فاز آموزش و آزمون

جدول ۳- تحلیل نتایج مدل‌های پیش‌بینی دبی جریان بر اساس معیارهای آماری در فازهای آموزش و آزمون

Table 3- Analysis of the results of streamflow prediction models based on statistical criteria in the training and testing phases

Method	Train				Test			
	R^2	RMSE	MAE	KGE	R^2	RMSE	MAE	KGE
RF-RFECV	0.954	85.476	31.942	0.816	0.675	155.673	77.665	0.574
RF-SFM	0.953	87.276	23.252	0.814	0.926	74.067	43.967	0.934
SVR-SFM	0.109	378.169	80.035	-0.092	0.182	246.922	78.196	0.048
LASSO	0.962	78.300	38.695	0.966	0.879	94.876	46.402	0.765
XGBoost	1	4.737	3.517	0.998	0.813	118.039	47.988	0.629
Climatology	0.048	390.771	102.245	-0.195	-0.017	275.388	106.606	-0.216

عملکرد مدل SVR-SFM

طبق جدول (۳) مدل SVR-SFM ضعیف‌ترین عملکرد را در میان مدل‌های یادگیری ماشین ارائه داده است. مقادیر بسیار پایین R^2=۰/۱۸۲ و KGE=۰/۰۴۸ در فاز آزمون، همراه با خطای بالا (RMSE=۲۴۶/۹۲۲)، پراکندگی گسترده نقاط در شکل (۳) و فاصله زیاد آن‌ها از خط ایده‌آل نشان‌دهنده ناتوانی این مدل در بازنمایی روابط غیرخطی پیچیده بین متغیرهای هیدرولوژیکی و دبی جریان است. فاصله زیاد نقاط از خط ایده‌آل y=x، کاهش توان پیش‌بینی مدل را تأیید می‌کند. عملکرد ضعیف مدل SVR را می‌توان به سه عامل اصلی نسبت داد:

- ماهیت حدی و بسیار غیرخطی پدیده پیش‌بینی جریان رودخانه
- حساسیت SVR به انتخاب پارامترها و مقیاس داده‌ها
- محدودیت هسته RBF در بازنمایی ترکیب پیچیده اثرات بارش، رطوبت خاک و تأخیرهای زمانی

این نتایج بیانگر آن است که SVR علی‌رغم کارایی در مسائل با ابعاد کوچک‌تر، برای مجموعه داده‌های دارای رفتار غیرخطی شدید و نوسانات بالا، عملکرد محدودی دارد و در مسائل

هیدرولوژیکی با تعداد ویژگی بالا و رفتار حدی، اغلب عملکرد ضعیف‌تری نسبت به مدل‌های تجمیعی دارد. با توجه به حساسیت SVR به پارامتر عرض کرنل (γ)، برای این پارامتر از مقدار پیش‌فرض استفاده شد که به‌طور خودکار بر اساس ویژگی‌های آماری داده‌های آموزشی تعیین می‌گردد. این رویکرد به افزایش قابلیت تعمیم مدل و کاهش خطر بیش‌برازش در مقایسه با انتخاب مقادیر ثابت γ به‌کار گرفته شد. سایر ابرپارامترهای مدل مانند C و ε نیز با استفاده از آزمون و خطا به‌ترتیب ۱ و ۰/۱ در نظر گرفته شد. استفاده از SFM قبل از آموزش، ابعاد فضای ویژگی را کاهش داده و حساسیت مدل به هم‌خطی متغیرها را کم و پایداری عددی آن را افزایش داده است.

عملکرد مدل LASSO

نتایج جدول (۳) نشان می‌دهد که مدل LASSO عملکرد نسبتاً پایداری در پیش‌بینی دبی جریان ارائه کرده است. مقادیر R^2=۰/۸۷۹ و KGE=۰/۷۶۵ در فاز آزمون نشان می‌دهد که این مدل توانسته الگوی کلی تغییرات دبی را تا حد مناسبی بازتولید کند. با این حال، مقادیر بالاتر RMSE=۹۴/۸۷۶ و MAE=۴۶/۴۰۲ نسبت به مدل RF-SFM بیانگر محدودیت این مدل در شبیه‌سازی رفتارهای غیرخطی و تغییرات شدید دبی جریان است. این نتایج نشان می‌دهد که با وجود توانایی LASSO در کاهش پیچیدگی مدل و کنترل بیش‌برازش، ماهیت خطی و ساده شده آن باعث می‌شود برخی از الگوهای غیرخطی جریان به‌درستی مدل‌سازی نشوند. بنابراین LASSO گزینه‌ای مناسب برای تحلیل‌های ساده و تفسیرپذیر است اما دقت آن نسبت به مدل‌های غیرخطی تجمیعی و انتخاب ویژگی توکار، کمتر است. همان‌طور که در شکل (۳) نشان داده شده است نمودار پراکنش مقادیر محاسباتی و مشاهداتی LASSOبیانگر تمرکز نقاط حول خط y=x است که با مقادیر R^2و KGE جدول (۳) همخوانی دارد. ضریب جریمه برابر ۰/۰۱ در نظر گرفته شد که با توجه به مقادیر پیش‌فرض و بررسی‌های اولیه مدل انتخاب گردید. نتایج، عملکرد پایدار اما محدود LASSOرا ارائه می‌کند و نشان می‌دهد که این مدل برای پیش‌بینی عملیاتی دبی جریان ماهانه، گزینه‌ای ساده و قابل تفسیر است اما برای بازنمایی دقیق تغییرات شدید و غیرخطی جریان مناسب‌ترین انتخاب نیست.

عملکرد مدل XGBoost

همان‌طور که در جدول (۳) نشان داده شده است مدل XGBoost بالاترین میزان برازش در فاز آموزش را داشته است (R^2=۱، KGE=۰/۹۹۸) که بیانگر توان بسیار بالای این الگوریتم در یادگیری ساختار داده‌هاست. با این حال، کاهش عملکرد در فاز آزمون (R^2=۰/۸۱۳، KGE=۰/۶۲۹) و افزایش خطاها (RMSE=۱۱۸/۰۳۹ و MAE=۴۷/۹۸۸) نشان‌دهنده بروز بیش‌برازش می‌باشد. اختلاف قابل توجه بین عملکرد داده‌های آموزش و آزمون نشان می‌دهد که پیچیدگی بالای XGBoost و حساسیت آن به تنظیم پارامترها می‌تواند در مجموعه داده‌های هیدرولوژیکی محدود باعث کاهش قابلیت تعمیم شود. این نتیجه تأکید می‌کند که حتی الگوریتم‌های قدرتمند نیز در مجموعه داده‌های هیدرولوژیکی محدود و پرنوسان، مستعد بیش‌برازش هستند. همان‌طور که در شکل (۳) نشان داده شده است نمودار پراکندگی مقادیر محاسباتی و مشاهداتی XGBoost نشان می‌دهد که در فاز آموزش نقاط به‌خوبی حول خط y=x متمرکز شده‌اند، اما در فاز آزمون پراکندگی نقاط افزایش یافته و برخی مقادیر بالای دبی، به‌صورت کم برآورد شده‌اند که با کاهش KGE و افزایش RMSE و MAE در جدول (۳) هم‌خوانی دارد. پارامترهای مدل شامل تعداد درخت‌ها برابر ۱۰۰ (پیش‌فرض)، عمق بیشینه هر درخت برابر شش، نرخ یادگیری ۰/۱ و جریمه وزن برگ‌ها (λ) برابر یک در نظر گرفته شد. پارامترجریمه برگ‌ها (γ) نیز برابر صفر در نظر گرفته شد تا هیچ محدودیتی برای ایجاد برگ‌های جدید اعمال نشود و مدل توانایی کامل در یادگیری ساختار داده‌ها در فاز آموزش داشته باشد. این مقادیر بر اساس ترکیبی از تنظیمات پیش‌فرض و آزمون و خطا انتخاب شده‌اند تا تعادل مناسبی بین دقت آموزش و مدیریت بیش‌برازش ایجاد شود. این تنظیمات پیچیده توانسته‌اند دقت بالایی در آموزش ایجاد کنند با وجود دقت بالای آموزش، محدودیت مجموعه داده باعث شد مدل در فاز آزمون همان عملکرد آموزش را حفظ نکند. این نتایج، نشان دهنده عملکرد قدرتمند اما حساس XGBoost است و تأکید می‌کند که این مدل برای پیش‌بینی عملیاتی دبی جریان ماهانه، علی‌رغم دقت بالا در آموزش، نیازمند توجه ویژه در تنظیم ابرپارامترها و مدیریت بیش‌برازش است.

مدل کلیماتولوژی

مدل کلیماتولوژی ضعیف‌ترین عملکرد را در میان تمامی مدل‌ها نشان داد. همان‌طور که در جدول (۳) نشان داده شده است مقادیر R^2=-۰/۰۱۷، RMSE=۲۷۵/۳۸۸، MAE=۱۰۶/۶۰۶ و KGE=-۰/۲۱۶ در فاز آزمون نشان‌دهنده ناتوانی کامل این مدل در بازتولید دامنه و نوسانات دبی جریان است. این نتیجه تأیید می‌کند که مدل‌های یادگیری ماشین اطلاعات معناداری فراتر از ساختار زمانی ساده داده‌ها استخراج می‌کنند. همان‌طور که در شکل (۳) نشان داده شده است نمودار پراکندگی نقاط مدل کلیماتولوژی تقریباً افقی است و هیچ هم‌خوانی معناداری با خط y=x ندارد که ضعف این مدل در شبیه‌سازی دبی جریان را به‌وضوح نشان می‌دهد. این نتایج تأیید می‌کند که این مدل تنها بر اساس میانگین بلندمدت دبی در هر ماه عمل می‌کند و بنابراین قادر به بازنمایی نوسانات بین‌سالی و رفتارهای حدی جریان نیست و برای پیش‌بینی جریان در حوضه‌های کوهستانی و با دینامیک پیچیده، نامعتبر است. با توجه به نتایج جدول (۳) و شکل (۳)، مدل کلیماتولوژی صرفاً به‌عنوان خط مبنا برای مقایسه عملکرد سایر مدل‌های یادگیری ماشین مورد

استفاده قرار گرفت. تفاوت واضح بین این مدل و سایر مدل‌های پیشرفته، نشان می‌دهد که الگوریتم‌های یادگیری ماشین توانایی استخراج الگوهای غیرخطی و پیچیده داده‌ها را دارند و در پیش‌بینی دبی ماهانه کارایی بالاتری ارائه می‌کنند. این مقایسه به‌طور شفاف برتری ذاتی یادگیری ماشین نسبت به مدل‌های ساده زمانی را نشان می‌دهد و استفاده از این مدل را صرفاً به‌عنوان خط مبنا توجیه می‌کند. با وجود آنکه روش کلیماتولوژی به‌عنوان یک روش پایه ساده برای ارزیابی اولیه عملکرد مدل‌ها مورد استفاده قرار گرفت، مقایسه مدل‌های یادگیری ماشین با روش‌های آماری متداول‌تر نظیر مدل‌های سری زمانی می‌تواند در مطالعه‌های آینده به درک جامع‌تری از کارایی این مدل‌ها در پیش‌بینی دبی رودخانه منجر شود.

لازم به ذکر است که توانایی بالای مدل‌های یادگیری ماشین در استخراج روابط پیچیده و غیرخطی در داده‌های هیدرولوژیکی، مدل‌های با ساختار پیچیده در صورت عدم کنترل مناسب ممکن است مستعد بیش‌برازش باشند. در چنین شرایطی، مدل ممکن است الگوهای موجود در داده‌های آموزش را به‌خوبی یاد بگیرد اما در تعمیم به داده‌های جدید عملکرد ضعیف‌تری نشان دهد. در این پژوهش به‌منظور ارزیابی توانایی تعمیم مدل‌ها، داده‌ها به دو بخش آموزش و آزمون تقسیم شدند و عملکرد مدل‌ها بر روی مجموعه داده آزمون مورد ارزیابی قرار گرفت. نتایج نشان داد که اگرچه برخی مدل‌ها تمایل به بیش‌برازش داشتند، اما مدل Random Forest توانست عملکرد نسبتاً پایدار و قابل قبولی در مجموعه داده آزمون ارائه دهد. با این حال، بررسی مدل‌های ساده‌تر یا رویکردهای ترکیبی می‌تواند در مطالعه‌های آینده برای دست‌یابی به تعادل بهتر میان دقت پیش‌بینی و سادگی مدل مورد توجه قرار گیرد.

مقایسه SFM و RFECV در انتخاب ویژگی‌ها

نتایج این مطالعه به‌وضوح نشان داد که دو رویکرد انتخاب ویژگی، منجر به مدل‌سازی‌های کاملاً متفاوتی می‌شوند. روش SFM با حفظ تعداد بیشتری از متغیرهای مرتبط، بیشترین دقت پیش‌بینی را ارائه داد. این موضوع به‌ویژه در عملکرد مدل RF-SFM مشهود است که بالاترین مقدار R^2=۰/۹۲۶ در مرحله آزمون را ثبت کرد (جدول ۳). در مقابل با حذف اکثر ویژگی‌های کم اهمیت و نگه داشتن تنها زیرمجموعه‌ای از مهم‌ترین متغیرها، RFECV مدلی ساده و پایدار ایجاد می‌کند. این رویکرد به سمت مدل حداقلی و باثبات‌تر حرکت کرده که کاهش قابل توجه دقت پیش‌بینی در فاز آزمون (R^2=۰/۶۷۵) را توجیه می‌کند. این تفاوت رفتاری نشان می‌دهد که RFECV بیشتر برای مطالعه‌هایی مناسب است که تفسیرپذیری، کاهش بعد و پایداری ساختار مدل در اولویت قرار دارد در حالی‌که SFM انتخاب مناسبی برای دست‌یابی به حداکثر دقت پیش‌بینی عملیاتی است. این نتایج نشان می‌دهد که انتخاب مدل باید بر اساس هدف نهایی انجام شود: برای پیش‌بینی عملیاتی با بیشترین دقت، RF-SFM، برای مدل‌سازی پایدار و کم‌ویژگی در تحلیل‌های علمی RF-RFECV، برای تحلیل‌های خطی و ساده، LASSO انتخاب می‌شود و برای مدل‌های بسیار منعطف با ریسک بالای بیش‌برازش، XGBoost به‌کار گرفته می‌شود. این مقایسه به‌وضوح نشان می‌دهد که در مدل‌سازی دبی جریان، یک مبادله بنیادی میان دقت پیش‌بینی، پایداری و تفسیرپذیری وجود دارد و انتخاب رویکرد مناسب وابسته به اهداف پژوهش است.

دیاگرام تیلور

به‌منظور ارزیابی هم‌زمان سه جنبه اصلی عملکرد مدل‌ها شامل همبستگی، واریانس نسبی و خطای کلی از نمودار تیلور استفاده شد. این نمودار امکان مقایسه بصری مدل‌ها را بر اساس سه شاخص آماری شامل ضریب همبستگی پیرسون، انحراف معیار نرمال شده و خطای میانگین مربعات فراهم می‌کند و به‌عنوان یکی از ابزارهای استاندارد در ارزیابی مدل‌های هیدرولوژیکی شناخته می‌شود (Taylor, 2001). در این نمودار، نزدیکی هر نقطه به موقعیت داده مرجع بیانگر عملکرد بهتر مدل از نظر بازتولید الگوی تغییرات دبی جریان است (Gleckler et al., 2008). همان‌طور که در شکل (۴) نشان داده شده است مدل RF-SFM نزدیک‌ترین موقعیت را به نقطه مرجع دارد و بیشترین همبستگی و انحراف معیار مشابه با مشاهدات را داراست. این امر توانایی بالای این مدل در بازنمایی هم‌زمان دامنه نوسانات و ساختار زمانی دبی جریان را نشان می‌دهد و با نتایج عددی جدول (۳) مبنی بر مقادیر بالای R^2 و KGEسازگاری است. در مقابل، مدل RF-RFECV اگرچه همبستگی قابل قبولی نشان می‌دهد، اما فاصله بیشتری از نقطه مرجع دارد که منعکس کننده کاهش بخشی از اطلاعات مؤثر به‌دلیل حذف تهاجمی ویژگی‌هاست؛ رفتاری که در مطالعه‌های پیشین نیز برای روش‌های حداقلی گزارش شده است (Kuhn and Johnson, 2013). مدل XGBoost در نمودار دارای انحراف معیار بزرگ‌تر از واحد و فاصله محسوس از مرجع است که نشان‌دهنده واریانس بیش از حد و احتمال بیش‌برازش می‌باشد؛ مسئله‌ای که در مدل‌های گرادیان تقویتی در صورت تنظیم نامناسب ابرپارامترها رایج است (Chen and Guestrin, 2016). مدل LASSO موقعیتی میانی دارد و اگرچه همبستگی مناسبی ایجاد کرده است اما به‌دلیل ماهیت خطی قادر به بازنمایی کامل رفتارهای غیرحدی دبی نبوده است (Tibshirani, 1996). ضعیف‌ترین عملکرد مربوط به مدل SVR-SFM و مدل پایه کلیماتولوژی است که با همبستگی پایین و انحراف معیار نامتناسب، ناتوانی آن‌ها را در شبیه‌سازی روابط غیرخطی پیچیده حوضه تأیید می‌کند. به‌طور کلی، تحلیل نمودار تیلور تأیید می‌کند که انتخاب روش انتخاب ویژگی نقش تعیین‌کننده‌ای در موقعیت مدل‌ها دارد؛ به‌طوری‌که رویکرد SFM با حفظ اطلاعات بیشتر، به نقطه مرجع نزدیک‌تر شده و برای پیش‌بینی عملیاتی مناسب‌تر است، در حالی‌که RFECV به سمت مدلی ساده‌تر اما با فاصله بیشتر از مرجع حرکت کرده است. این یافته با اصل مبادله میان دقت و

تفسیرپذیری در مدل‌سازی هیدرولوژیکی همخوانی دارد (Molnar, 2020).

نمودار سری زمانی شکل (۵) نیز مقایسه دبی مشاهده شده و پیش‌بینی شده را در ایستگاه کشکان-پلدختر نشان می‌دهد. مدل RF-SFM توانسته است الگوی زمانی و روندهای فصلی جریان را با دقت مناسبی بازتولید کند و در اغلب بازه‌ها تغییرات افزایشی و کاهشی دبی را دنبال نماید. تطابق مناسب مقادیر محاسباتی با مشاهداتی، نشان‌دهنده توانایی مدل در استخراج روابط حاکم بر رژیم جریان و بهره‌گیری مؤثر از متغیرهای ورودی است. این توانایی پیش‌بینی دقیق، می‌تواند برای برنامه‌ریزی و مدیریت منابع آب در حوضه‌های رودخانه‌ای با رژیم متغیر، بسیار ارزشمند باشد.

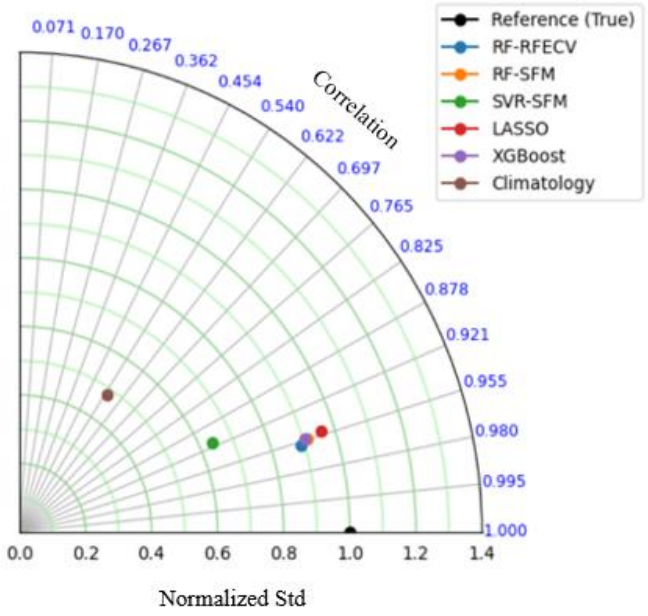


Fig. 4- Taylor diagram for comparing the performance of streamflow prediction models

شکل ٤- دیاگرام تیلور برای مقایسه عملکرد مدل‌های پیش‌بینی دبی جریان

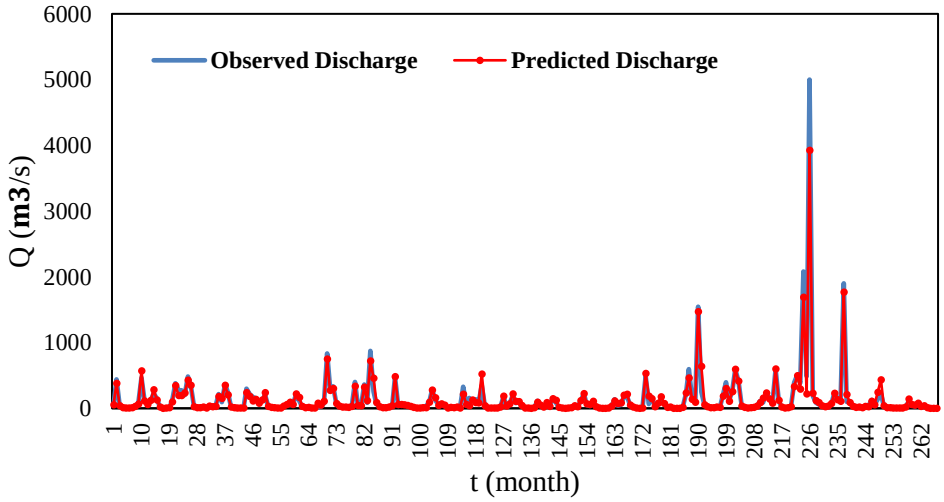


Fig. 5- Comparison of the time series of observed and predicted streamflow by the RF-SFM model in the testing phase

شکل ٥- مقایسه سری زمانی دبی مشاهده شده و پیش‌بینی شده توسط مدل RF-SFM در فاز آزمون

نتیجه‌گیری

در این پژوهش، عملکرد چندین روش یادگیری ماشین شامل جنگل تصادفی، XGBoost، رگرسیون LASSO و ماشین بردار پشتیبان برای پیش‌بینی دبی ماهانه رودخانه کشکان–پلدختر مورد ارزیابی قرار گرفت. به‌منظور بررسی نقش انتخاب ویژگی در بهبود دقت، پایداری و قابلیت تعمیم مدل‌ها، دو رویکرد انتخاب ویژگی SFM و RFECV به‌کار گرفته شد و نتایج با یک مدل پایه کلیماتولوژی مقایسه گردید. نتایج نشان داد که تمامی مدل‌های مبتنی بر یادگیری ماشین عملکردی به‌مراتب بهتر از مدل پایه کلیماتولوژی دارند که بیانگر توان بالای روش‌های داده‌محور در استخراج الگوهای پیچیده و غیرخطی حاکم بر جریان رودخانه است. در میان مدل‌ها، مدل RF-SFM بهترین عملکرد کلی را ارائه داد و با دست‌یابی به مقادیر بالای R^2 و KGE در فاز آزمون توانست تعادلی مطلوب میان دقت پیش‌بینی و قابلیت تعمیم برقرار کند. این موضوع نشان می‌دهد که حفظ مجموعه‌ای غنی از ویژگی‌های مؤثر، در کنار ساختار تجمیعی جنگل تصادفی، نقش کلیدی در بازنمایی دینامیک پیچیده دبی جریان دارد. در مقابل، RF-RFECV با وجود کاهش دقت پیش‌بینی در فاز آزمون، ساختاری ساده‌تر و پایدارتر ارائه داد به‌گونه‌ای که RFECV با حذف متغیرهای هم‌بسته و کم‌اهمیت، به سمت مدلی حداقلی و با قابلیت تعمیم بالاتر حرکت می‌کند. از این رو، انتخاب میان SFM و RFECVبه هدف نهایی مطالعه وابسته است؛ به‌طوری‌که RFECV برای تحلیل‌های تفسیرپذیر و کم ویژگی و SFM برای کاربردهای عملیاتی با حداکثر دقت مناسب‌تر است. مدل LASSO عملکردی نسبتاً پایدار و قابل قبول از خود نشان داد، اما ماهیت خطی آن موجب محدودیت در شبیه‌سازی رفتارهای غیرخطی و تغییرات شدید دبی جریان شد.

مدل XGBoost اگرچه در فاز آموزش برازش تقریباً کاملی ارائه داد، اما افت عملکرد در فاز آزمون نشان‌دهنده حساسیت بالای این الگوریتم به تنظیم ابرپارامترها و بروز بیش‌برازش در مجموعه داده‌های هیدرولوژیکی محدود است. همچنین عملکرد ضعیف مدل-SVR SFMنیز نشان داد که این روش در مواجهه با داده‌های دارای رفتار غیرخطی شدید و نوسانات بالا، کارایی کمتری نسبت به مدل‌های تجمیعی مبتنی بر درخت تصمیم دارد. به‌طور کلی، نتایج این مطالعه نشان می‌دهد که هیچ مدل واحدی به‌طور مطلق برتر از سایر روش‌ها نیست و عملکرد نهایی مدل‌ها به ترکیبی از نوع الگوریتم، راهبرد انتخاب ویژگی و هدف کاربردی وابسته است. یافته‌های این پژوهش بر اهمیت انتخاب آگاهانه ساختار مدل‌سازی و روش انتخاب ویژگی در پیش‌بینی دبی جریان رودخانه تأکید داشته و می‌تواند به‌عنوان راهنمایی عملی برای توسعه مدل‌های پیش‌بینی دقیق، پایدار و قابل تعمیم در مدیریت و بهره‌برداری منابع آب مورد استفاده قرار گیرد. در پژوهش‌های آتی، به‌کارگیری مدل‌های یادگیری عمیق، ادغام داده‌های سنجش از دور و متغیرهای فیزیکی حوضه و توسعه رویکردهای ترکیبی انتخاب ویژگی می‌تواند به بهبود دقت و قابلیت تعمیم مدل‌های پیش‌بینی دبی جریان منجر شود. همچنین ارزیابی پایداری مدل‌ها تحت سناریوهای اقلیمی و زمانی متفاوت، درک جامع‌تری از عملکرد آن‌ها در سامانه‌های هیدرولوژیکی فراهم خواهد آورد.

تضاد منافع نویسندگان

نویسندگان این مقاله اعلام می‌دارند که هیچ گونه تضاد منافعی در خصوص نگارش و انتشار مطالب و نتایج این پژوهش ندارند.

منابع مالی

نویسندگان هیچ گونه حمایت مالی برای تحقیق، تالیف و انتشار این مقاله دریافت نکردند.

دسترسی داده‌ها

مجموعه داده‌های تولید شده و یا تحلیل شده در طول مطالعه حاضر، در صورت درخواست معقول، از نویسنده مسئول در دسترس هستند.

مشارکت نویسندگان

نویسنده اول: مریم میربیک سبزواری، جمع آوری داده‌ها، تجزیه و تحلیل داده‌ها، نگارش مقاله.

نویسنده دوم: حسن ترابی‌پوده، استاد راهنمای اول، تجزیه و تحلیل داده‌ها.

نویسنده سوم: محمد باقر دولتشاهی، استاد مشاور دوم، بررسی نتایج.

نویسنده چهارم: امیر حمزه حقی‌آبی، استاد راهنمای دوم، اصلاح مقاله.

نویسنده پنجم: عباس ملکی، استاد مشاور اول، بررسی مقاله.

تشکر و قدردانی

به‌این وسیله نویسندگان مقاله مراتب تقدیر و تشکر خود را از شرکت آب منطقه‌ای استان لرستان به‌دلیل در اختیار قرار دادن آمار و اطلاعات مورد استفاده در این پژوهش ابراز می‌نمایند.

References

1- Breiman, L., 2001. Random forests. *Machine Learning*, 45(1), pp.5–32. https://doi.org/10.1023/A:1010933404324.

2- Chen, T. and Guestrin, C., 2016. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD*, pp.785–794. https://doi.org/10.1145/2939672.2939785

3- Chu, H., Wei, J. and Qiu, J., 2018. Monthly streamflow forecasting using EEMD-Lasso-DBN method based on multi-scale predictors selection. *Water*, 10(10), 1486. https://doi.org/10.3390/w1010148.

4- Costa, A. C., Estacio, A. B. S., de Souza Filho, F. A. and Lima Neto, I. E., 2021. Monthly and seasonal streamflow forecasting of large dryland catchments in Brazil. *Journal of Arid Land*, 13(3), pp. 205–223. https://doi.org/ 10.1007/s40333-021-0062-4.

5- Crochemore, L., Ramos, M.H., Pappenberger, F. and Perrin, C., 2017. Seasonal streamflow forecasting by conditioning climatology with precipitation indices. *Hydrology and Earth System Sciences*, 21, pp.1573–1591. https://doi.org/10.5194/hess-21-1573-2017.

6- Elhallaoui Oueldkaddour, F. Z., Wariaghli, F., Brirhet, H., Yahyaoui, A., and Jaziri, H., 2025. Comparison of machine learning models for real time flow forecasting in the semi-arid Bouregreg basin. *Limnological Review*, pp. 25(1), 6. https://doi.org/10.3390/limnolrev25010006.

7- Fang, W., Ren, K., Liu, T., Shang, J., Jia, S., Jiang, X. and Zhang, J., 2024. An evaluation of random forest based input variable selection methods for one month ahead streamflow forecasting. *Scientific Reports*, 14, Article 29766. https://doi.org/10.1038/s41598-024-81502-y.

8- Fathian, F., Mehdizadeh, S., Sales, A. K. and Safari, M. J. S, 2019. Hybrid models to improve the monthly river flow prediction: Integrating artificial intelligence and nonlinear time series models. *Journal of Hydrology*, 575, pp.1200–1213. https://doi.org/10.1016/j.jhydrol.2019.06.025.

9- Ghorbani, MA., R. Khatibi, A. Geol, M.H. Fazelifard and Azani, A., 2016. Modeling river discharge time series using support vector machine and artificial neural networks. *Environmental Earth Sciences*, 75(4): pp.675-685. https://doi.org/10.1007/s12665-016-5435-6.

10- Gleckler, P.J., Taylor, K.E. and Doutriaux, C., 2008. Performance metrics for climate models. Journal of Geophysical Research, 113, D06104. https://doi.org/10.1029/2007JD008972.

11- Gupta, H.V., Kling, H., Yilmaz, K.K. and Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. Journal of Hydrology, 377(1–2), pp.80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003.

12- Guyon, I. and Elisseeff, A., 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, pp.1157–1182.

13- Guyon, I., Weston, J., Barnhill, S. and Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1–3), pp.389–422. https://doi.org/10.1023/A:1012487302797

14- Hao, R. and Bai, Z., 2023. Comparative study for daily streamflow simulation with different machine learning methods. *Water*, 15(6), 1179. https://doi.org/10.3390/w15061179.

15- Hastie, T., Tibshirani, R. and Friedman, J, 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). *Springer*, New York. https://link.springer.com/book/10.1007/978-0-387-84858-7.

16- Ikram, R.M.A.; Goliatt, L.; Kisi, O.; Trajkovic, S.; Shahid, S. Covariance matrix adaptation evolution strategy for improving machine learning approaches in streamflow prediction. Mathematics 2022, 10, 2971. https://doi.org/10.3390/math10162971.

17- Kisi, O., 2015. River flow modeling using artificial Neural Networks. *Journal of Hydrology*, 530, pp.421–436. https://doi.org/10.1061/(ASCE)1084-0699(2004)9:1(60).

18- Ko, C.M., Jeong, Y.Y., Lee, Y.M. and Kim, B.S, 2020. The development of a quantitative precipitation forecast correction technique based on machine learning for hydrological applications. *Atmosphere*, 11(1), 111. https://doi.org/10.3390/atmos11010111.

19- Kuhn, M. and Johnson, K., 2013. Applied predictive modeling. Springer. https://doi.org/10.1007/978-1-4614-6849-3.

20- Kwon, S., Lee, S. and Na, O, 2017. Tuning parameter selection for the adaptive Lasso in the autoregressive model. *J. Korean Stat. Soc*, 46, pp. 285–297. https://doi.org/10.1016/j.jkss.2016.10.005

21- Liaw, A. and Wiener, M, 2002. Classification and regression by random Forest. *R News*, 2(3), pp. 18–22. https://journal.r-project.org/articles/RN-2002-022.

22- Liu, J., Ren, K., Ming, T., Qu, J., Guo, W. and Li, H, 2022. Investigating the effects of local weather, streamflow lag, and global climate information on 1-month-ahead streamflow forecasting by using XGBoost and SHAP: Two case studies involving the contiguous USA. *Acta Geophys*, 71, pp. 905–925. https://doi.org/10.1007/s11600-022-00928-y.

23- Liu, Z., Zhou, P., Chen, G. and Guo, L. 2014. Evaluating a coupled discrete wavelet transform and support vector regression for daily and monthly streamflow forecasting. *Journal of Hydrology*, 519, pp. 2822–2831. https://doi.org/10.1016/j.jhydrol.2014.06.050.

24- Lorestan Regional Water Company, 2014. Hydrological and meteorological data of Lorestan Province, Lorestan, Iran.

25- Ma, W., Zhang, X., Shen, Y., Xie, J., Zuo, G., Zhang, X., and Jin, T., 2024. Incorporating recursive feature elimination and decomposed ensemble modeling for monthly runoff prediction. *Water*, 16(21), 3102. https://doi.org/10.3390/w16213102. https://doi.org/10.3390/w16213102.

26- Maier, H.R., Jain, A., Dandy, G.C., and Sudheer, K.P, 2010. Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. *Environmental Modelling & Software*, 25 (8), 891–909. https://doi.org/10.1016/j.envsoft.2010.02.003.

27- Marco, J.B., Harboe, R. and Salas, J.D. (Eds.), 1993. Stochastic hydrology and its use in water resources systems simulation and optimization.

28- Molnar, C., 2020. Interpretable machine learning: A guide for making black box models explainable. *Lulu Press*.

29- Mosavi, A., Ozturk, P., and Chau, K. W., 2018. Flood prediction using machine learning models. *Water*, 10(11), 1536. https://doi.org/10.3390/w10111536.

30- Nagelkerke, N.J.D, 1991. A note on a general definition of the coefficient of determination. *Biometrika*, 78(3), pp. 691-692. https://doi.org/10.1093/biomet/78.3.691.

31- Nearing, G.S., Kratzert, F., Klotz, D., Gauch, M., Sampson, A.K., and Hochreiter, S., 2021. What role does Hydrological Science Play in the Age of Machine Learning? *Water Resources Research*, 57(3). https://doi.org/10.1029/2020WR028091

32- Ni, L., Wang, D., Wu, J., Wang, Y., Tao, Y., Zhang, J., and Liu, J., 2020. Streamflow forecasting using extreme gradient boosting model coupled with Gaussian mixture model. *Journal of Hydrology*, 586, 124901. https://doi.org/10.1016/j.jhydrol.2020.124901.

33- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É., 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, pp. 2825–2830.

34- Potdar, A.S., Kirstetter, P., Woods, D. and Saharia, M, 2021. Toward predicting flood event peak discharge in ungauged basins by learning universal hydrological behaviors with machine learning. *J. Hydrometeorol*, 22, pp. 2971–2982. https://doi.org/10.1175/JHM-D-20-0302.1.

35- Shen, C., 2018. A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resources Research*, 54(11), 8558–8593. https://doi.org/10.1029/2018WR022643.

36- Smola, A.J. and Schölkopf, B.A., 2004. A tutorial on support vector regression. *Stat. Comput*. 14, pp. 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88.

37- Szczepanek, R., 2022. Daily streamflow forecasting in mountainous catchment using XGBoost, LightGBM and CatBoost. *Hydrology*, 9(12), 226. https://doi.org/10.3390/hydrology9120226.

38- Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres*, *106*(D7), pp. 7183–7192. https://doi.org/10.1029/2000JD900719

39- Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

40- Tongal, H., and Booij, M.J., 2018. Simulation and forecasting of streamflows using machine learning models coupled with base flow separation. *Journal of Hydrology*, 564, pp. 266–282. https://doi.org/10.1016/j.jhydrol.2018.07.004.

41- Tyralis, H., Papacharalampous, G. and Langousis, A., 2019. A brief review of random forests for water scientists and practitioners and their recent history in water resources. *Water*, 11(5), 910. https://doi.org/10.3390/w11050910

42- Vapnik, V., 1995. The Nature of Statistical Learning Theory; Springer: New York, NY, USA.

43- Yaseen, Z.M., El-Shafie, A., Jaafar, O., Afan, H.A. and Sayl, K.N., 2015. Artificial intelligence based models for stream-flow forecasting: 2000–2015. *Journal of Hydrology*, 530, 829–844. https://doi.org/10.1016/j.jhydrol.2015.10.038.

44- Yu, P.S., Chen, S.T. and Chang, I.F., 2006. Support vector regression for real time flood stage forecasting. *Journal of Hydrology*, 328(3–4), pp. 704–716. https://doi.org/10.1016/j.jhydrol.2006.01.021.

45- Zhang, X., Li, H., Wang, Y. and Chen, J., 2026. SFM_MF: A streamflow forecasting model based on model fusion for small sample data in small and medium-sized rivers. *Applied Intelligence*. 56(2), 60. https://doi.org/10.1007/s10489-025-07074-0.

46- Zounemat-Kermani, M., Batelaan, O., Fadaee, M., and Hinkelmann, R., 2021. Ensemble Machine learning paradigms in hydrology: A review. *Journal of Hydrology*, 598, 12626. https://doi.org/10.1016/j.jhydrol.2021.126266.

پذیرش نهایی- قبل از چاپ

EXTENDED ABSTRACT

Evaluation of the Impact of Different Feature Selection Approaches on River Flow Modeling Using Machine Learning Algorithms

M. Mirbeyk Sabzevari[1], H. Torabi Podeh[2*], M. Bagher Dowlatshahi[3], A. H. Haghiabi[4] and A. Maleki[5]

1. Ph.D. Candidate in Hydraulic Structures, Department of Water Engineering, Faculty of Agriculture, Lorestan University.
2*. Professor, Department of Water Engineering, Faculty of Agriculture, Lorestan University. torabi.ha@lu.ac.ir
3. Associate Professor, Department of Computer Engineering, Faculty of Engineering, Lorestan University.
4. Professor, Department of Water Engineering, Faculty of Agriculture, Lorestan University.
5. Associate Professor, Department of Water Engineering, Faculty of Agriculture, Lorestan University.

ARTICLE INFO

Artlcle history:
Received:
Revised:
Accepted:
Available online:

Keywords:
Flow Prediction, Feature Selection, Random Forest, Machine Learning, Climatology Baseline Model.

TO CITE THIS ARTICLE:

Abstract

Accurate prediction of river flow remains a major challenge in hydrological systems due to nonlinear relationships and temporal variability governing streamflow behavior. Machine learning algorithms are increasingly used for this purpose because of their strong capability in modeling complex nonlinear interactions among hydrological variables. However, when a large number of input predictors are used, model complexity, computational cost, and the risk of overfitting increase, highlighting the importance of applying effective feature selection methods.

This study evaluates the impact of different feature selection strategies on the performance of machine learning models for monthly streamflow prediction of the Kashkan–Poldokhtar River. Two feature selection approaches SelectFromModel (SFM), based on model-derived feature importance, and recursive feature elimination with cross-validation (RFECV) were implemented alongside the inherent feature-handling mechanisms embedded within certain algorithms. Four machine learning models, including Random Forest (RF), Extreme Gradient Boosting (XGBoost), Least Absolute Shrinkage and Selection Operator (LASSO), and Support Vector Regression (SVR), were trained and tested. Their performances were evaluated using R^2, RMSE, MAE, and KGE metrics and compared with a climatology baseline. Results show that feature selection significantly influences the predictive capability of machine learning models. The RF–SFM model achieved the best testing performance ($R^2 = 0.926$, KGE = 0.934), demonstrating the effectiveness of SFM in identifying key predictors and improving forecast accuracy. The RF-RFECV model produced a simpler and more stable structure by iteratively removing less important features, making it suitable for interpretation and scientific analysis. XGBoost showed a tendency toward overfitting, while SVR exhibited the weakest performance on this dataset. Comparison with the climatology baseline confirmed the superiority of machine learning methods in representing complex hydrological dynamics.

Overall, selecting appropriate feature selection strategies is essential for balancing predictive accuracy and model simplicity and can improve operational streamflow forecasting and support water resources management.

Introduction

Accurate streamflow prediction is essential for water resources management, reservoir operation, and mitigation of flood and drought impacts. Traditional statistical models are limited by linear assumptions and data stationarity, reducing their ability to capture nonlinear behaviors and extreme events. In contrast, machine learning algorithms have gained widespread attention for streamflow

forecasting due to their ability to extract complex patterns and adapt to non-stationary processes, although their performance strongly depends on input data quality and feature selection (Mosavi et al., 2018; Shen, 2018). This study focuses on monthly streamflow prediction at the Kashkan-Poldokhtar River station in Iran, comparing two feature selection strategies, SelectFromModel (SFM), Recursive Feature Elimination with Cross-Validation (RFECV), along with an embedded feature selection approach within four machine learning algorithms RF, XGBoost, LASSO, and SVR to evaluate a balance between predictive accuracy and operational generalization.

Methodology

The Kashkan river catchment (7,950.2 km²) located in southwestern Iran, is a sub-basin of the Karkheh River. Monthly precipitation data from nine meteorological stations and streamflow data from two hydrometric stations were used. The input variables included current and lagged precipitation for up to three previous months P(t), P(t−1), P(t−2), P(t−3) and upstream streamflow at the Kakareza station, while the target variable was monthly streamflow at the Kashkan-Poldokhtar station. The study period spans from december 1969 to september 2024, resulting in 633 samples after data quality control. All preprocessing and modeling procedures were implemented in Python.

Five machine learning approaches were evaluated, including Random Forest combined with recursive feature elimination (RF-RFECV), Random Forest with SelectFromModel (RF-SFM), Support Vector Regression with feature selection (SVR-SFM), LASSO regression, and XGBoost. In addition, a climatology-based baseline model.

Random Forest (RF) is an ensemble tree-based algorithm used for modeling nonlinear relationships while minimizing overfitting (Breiman, 2001). The RF-RFECV model combines Random Forest with recursive feature elimination to select an optimal subset of predictors (Guyon et al., 2002). RF-SFM utilizes Random Forest for both prediction and embedded feature selection by prioritizing key variables (Molnar, 2020). In the SVR-SFM approach, feature importance is ranked to select the most relevant inputs for training the Support Vector Regression (SVR) model (Smola and Schölkopf, 2004). LASSO regression employs regularization to perform automatic feature selection and coefficient shrinkage (Kwon et al., 2017). XGBoost is a gradient-boosted decision tree algorithm optimized for high predictive accuracy and model complexity control (Chen and Guestrin, 2016). Finally, the climatology model uses historical monthly mean streamflow as a baseline to evaluate the performance gains of the machine learning models (Crochemore et al., 2017).

Results and Discusspn

Table (1) indicates that machine learning models significantly outperformed the baseline climatology model in predicting monthly discharge at the Kashkan-Poldokhtar station. Based on R^2 and KGE, the RF-SFM model showed the best performance, accurately capturing seasonal flow patterns ($R^2 = 0.926$, KGE = 0.934). The RF-RFECV model was simpler but less accurate ($R^2 = 0.675$). LASSO was stable but limited in representing nonlinear flow behavior ($R^2 = 0.879$). XGBoost overfitted the training data ($R^2 = 1$) and underperformed during testing ($R^2 = 0.813$). Both SVR-SFM and the climatology model exhibited the lowest accuracy. Overall, feature selection with SFM preserved informative variables and improved predictions, whereas RFECV produced simpler models with reduced accuracy. These results highlight the critical role of combining the choice of machine learning algorithm and feature selection method to achieve a balance between predictive accuracy and model generalization.

Table 1- Analysis of the results of streamflow prediction models based on statistical criteria in the training and testing phases

Method	Train				Test			
	R^2	RMSE	MAE	KGE	R^2	RMSE	MAE	KGE
RF-RFECV	0.954	85.476	31.942	0.816	0.675	155.673	77.665	0.574

RF-SFM	0.953	87.276	23.252	0.814	0.926	74.067	43.967	0.934
SVR-SFM	0.109	378.169	80.035	-0.092	0.182	246.922	78.196	0.048
LASSO	0.962	78.300	38.695	0.966	0.879	94.876	46.402	0.765
XGBoost	1	4.737	3.517	0.998	0.813	118.039	47.988	0.629
Climatology	0.048	390.771	102.245	-0.195	-0.017	275.388	106.606	-0.216

Conclusions

This study evaluated several machine learning models, including Random Forest, XGBoost, LASSO, and SVR, for monthly streamflow prediction in the Kashkan-Poldokhtar River. Two feature selection methods, SFM and RFECV, were also tested. All ML models outperformed the baseline climatology model. RF-SFM achieved the best balance between accuracy and generalization, while RF-RFECV produced a simpler and more stable model. LASSO and SVR showed limitations in capturing nonlinear flow patterns, and XGBoost was prone to overfitting. Overall, model choice and feature selection should align with the study's objectives, highlighting the importance of informed model design for accurate and robust streamflow prediction.

Conflict of Interest

The authors declared no potential conflicts of interest concerning the research, authorship, and publication of this article. Confirmation.

Funding

The authors received no financial support for the research, authorship, and publication of this article.

Data Availability Statements

The datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

Authors' contribution

First Author: Maryam Mirbeyk Sabzevari, Data collection, Data analysis, Writing draft.
Second Author: Hasan Torabi Podeh, First supervisor, Data analysis.
Third Author: Mohammad Bagher Dowlatshahi, Second advisor, Reviewing the results.
Fourth Author: Amir Hamzeh Haghiabi, Second supervisor, Manuscript revision.
Fifth Author: Abbas Maleki, First advisor, Manuscript review.

Acknowledgment

The authors would like to express their gratitude to the Lorestan Regional Water Company for providing the data and information used in this research.

References

1- Breiman, L., 2001. Random forests. *Machine Learning*, 45(1), pp.5–32. https://doi.org/10.1023/A:1010933404324.

2- Chen, T. and Guestrin, C., 2016. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD*, pp.785–794. https://doi.org/10.1145/2939672.2939785.

3- Crochemore, L., Ramos, M.H., Pappenberger, F. and Perrin, C., 2017. Seasonal streamflow forecasting by conditioning climatology with precipitation indices. *Hydrology and Earth System Sciences*, 21, pp.1573–1591. https://doi.org/10.5194/hess-21-1573-2017.

4- Guyon, I., Weston, J., Barnhill, S. and Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. *Machine Learning,* 46(1–3), pp.389–422. https://doi.org/10.1023/A:1012487302797

5- Kwon, S., Lee, S. and Na, O, 2017. Tuning parameter selection for the adaptive Lasso in the autoregressive model. *J. Korean Stat. Soc*, 46, pp. 285–297. https://doi.org/10.1016/j.jkss.2016.10.005.

6- Molnar, C., 2020. Interpretable machine learning: A guide for making black box models explainable. Lulu Press.

7- Mosavi, A., Ozturk, P., and Chau, K. W., 2018. Flood prediction using machine learning models. Water, 10(11), 1536. https://doi.org/10.3390/w10111536.

8- Shen, C., 2018. A transdisciplinary review of deep learning research and its relevance for water resources scientists. Water Resources Research, 54(11), 8558–8593. https://doi.org/10.1029/2018WR022643.

9- Smola, A.J. and Schölkopf, B.A., 2004. A tutorial on support vector regression. Stat. Comput. 14, pp. 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88.