نشریه آبیاری و زهکشی ایران
شماره۴، جلد ۱۷، مهر - آبان ۱۴۰۲، ص. ۷۰۱-۶۸۵

Iranian Journal of Irrigation and Drainage
No.4, Vol. 17, Oct.-Nov. 2023, p. 685-701

مقاله علمی-پژوهشی

توسعه مدل‌های تلفیقی یادگیری ماشین مبتنی بر روش تجزیه مد تجربی گروهی کامل در برآورد جریان ورودی به سد (مطالعه موردی سد دز)

نوید موسی زاده[۱]، علی محمد آخوندعلی[۲]، فرشاد احمدی[۳*]

تاریخ دریافت: ۱۴۰۱/۱۲/۱۶ تاریخ پذیرش: ۱۴۰۲/۰۲/۲۸

چکیده

برآورد جریان ورودی به مخزن سدها در برنامه‌ریزی و مدیریت بهینه منابع آب، تامین آب مورد نیاز بخش‌های مختلف و مدیریت سیلاب از اهمیت ویژه‌ای برخوردار است. بنابراین در پژوهش حاضر سعی گردید تا عملکرد مدل‌های یادگیری ماشین جنگل‌های تصادفی (RF) و رگرسیون فرآیند گاوسی (GPR) با استفاده از روش پیش‌پردازش داده‌های تجزیه مد تجربی گروهی کامل (CEEMD) در برآورد جریان ماهانه ورودی به سد دز در دوره آماری ۵۱-۱۳۵۰ تا ۹۷-۱۳۹۶ مورد بررسی قرار گیرد. بدین منظور الگوهای ورودی در چهار سناریوی مختف شامل استفاده از داده‌های جریان با تاخیرهای زمانی، ترکیب داده‌های جریان و بارش با تاخیرهای زمانی و اضافه کردن خاصیت تناوبی به دو حالت قبل آماده و به مدل‌های منفرد معرفی شدند. نتایج نشان داد که هر مدل با سناریوهای متفاوتی به حداکثر دقت خود دست می‌یابد و در این بین مدل GPR با شاخص RMSE برابر با ۹۷/۴۹ (m^3/s) بهترین عملکرد را داشت. پس از تعیین الگوهای برتر ورودی در هر سناریو، داده‌های مربوطه توسط روش CEEMD تجزیه و فرآیند مدل‌سازی با روش‌های RF و GPR انجام شد. بر اساس معیارهای ارزیابی، کاهش خطا و افزایش دقت در مدل‌های تلفیقی توسعه داده شده به طور قابل ملاحظه‌ای مشهود بود. به طوریکه مدل CEEMD-GPR تواتست مقدار شاخص RMSE را به طور متوسط حدود ۴۷ مترمکعب بر ثانیه کاهش دهد. همین روند برای مدل CEEMD-RF نیز مشاهده شد. به طور کلی عملکرد CEEMD-GPR در مقایسه با کلیه مدل‌های توسعه داده شده (منفرد یا تلفیقی) مناسب‌تر بوده و برای پیش‌بینی جریان ورودی به سد دز توصیه می‌شود.

واژه‌های کلیدی: تابع مد ذاتی، تاخیر، مدل تلفیقی، مدل منفرد

مقدمه

یکی از بخش‌های مهم در مدیریت منابع آب حوضه‌های آبریز، برآورد جریان رودخانه می‌باشد. این مهم در بهره‌برداری بهینه از مخازن سدها، توسعه سامانه‌های هشدار سیل و تولید نیروی برقابی از اهمیت ویژه‌ای برخوردار است. برای این منظور روش‌ها و مدل‌های متعدد ساده و پیچیده‌ای توسعه داده شده‌اند. به عنوان مثال انواع مدل‌های مفهومی بارش- رواناب، مدل‌های سری‌زمانی و الگوهای ترکیبی (هیبرید) ارائه شده است که می‌توان همه آنها را در زمره مدل‌های فیزیکی و مدل‌های داده محور طبقه‌بندی کرد (Lin et al., 2021). مدل‌های فیزیکی از توابع ریاضی برای بیان روابط موجود بین پدیده‌های مختلف هیدرولوژیکی استفاده نموده و به همین دلیل کاربرد آنها نیازمند مشاهده، تجربه و آزمایش‌های مختلف می‌باشد (Beven, 2020). به عبارت دیگر، در مدل‌های فیزیکی نتیجه نهایی مدل به شرایط مختلفی وابسته بوده و در نتیجه هزینه مدل‌سازی بسیار زیاد می‌باشد. علاوه بر این، بروز خطا به دلیل عدم قطعیت بالا در مدل‌های فیزیکی برای پیش‌بینی پارامتر پیچیده‌ای، همچون جریان رودخانه، اطمینان به این دسته از مدل‌ها را تحت تاثیر قرار می‌دهد (Lin et al., 2021). با توجه به مشکلات موجود در ساختار مدل‌های فیزیکی، محققان روش‌های مبتنی بر داده را در سال‌های اخیر توسعه داده و در پیش‌بینی‌های هیدرولوژیک استفاده کرده‌اند. از

۱ - دانشجوی کارشناسی ارشد مندسی منابع آب، دانشکده مهندسی آب و محیط زیست، دانشگاه شهید چمران اهواز، اهواز، ایران
۲ - استاد گروه هیدرولوژی و منابع آب، دانشکده مهندسی آب و محیط زیست، دانشگاه شهید چمران اهواز، اهواز، ایران
۳ - نویسنده مسئول و استادیار گروه هیدرولوژی و منابع آب، دانشکده مهندسی آب و محیط زیست، دانشگاه شهید چمران اهواز، اهواز، ایران
(*- نویسنده مسئول: (Email: f.ahmadi@scu.ac.ir
DOR: 20.1001.1.20087942.1402.17.4.7.7

جمله مزایای این روش‌ها می‌توان به اشکال‌زدایی آسان‌تر آنها، نیاز به تعداد متغیرهای ورودی کمتر و کاربرد فراگیر آنها اشاره نمود (Meng et al., 2021). روش‌های یادگیری ماشین در زمره مدل‌های داده-محور بوده و در توضیح روابط پیچیده عملکرد بسیار مناسبی داشته و به عنوان یک روش جایگزین موفق در پیش‌بینی جریان رودخانه پیشنهاد می‌شوند که از آن جمله می‌توان به مدل‌های رگرسیون فرآیند گاوسی[1] (GPR) و جنگل‌های تصادفی[2] (RF) اشاده نمود.

مدل فرآیندهای گاوسی یک چارچوب یادگیری ماشینی با نظارت احتمالی است که به طور گسترده برای حل مسائل رگرسیون غیرخطی و پیچیده استفاده می‌شود. یک مدل رگرسیون فرآیندهای گاوسی (GPR) می‌تواند پیش‌بینی‌هایی را با ترکیب دانش قبلی (هسته‌ها) انجام داده و معیارهای عدم قطعیت را در مورد آنها ارائه دهد (Wang, 2020). از این روش در مطالعات متعددی برای پیش بینی متغیرهای هیدرولوژیک استفاده شده است. ثاقبیان در پژوهشی از روش‌های مبتنی بر کرنل ماشین بردار پشتیبان (SVM) و رگرسیون فرآیند گاوسی (GPR) برای پیش‌بینی جریان روزانه رودخانه بهره برد. نتایج حاصل از تحلیل الگوهای تعریف‌شده نشان داد که مدل‌های SVM و GPR از توانایی لازم برای پیش‌بینی جریان کوتاه مدت رودخانه برخودار هستند (ثاقبیان، ۱۳۹۹). نیو و فنگ برای پیش‌بینی سری جریان روزانه ورودی به دو مخزن عظیم برق آبی در چین از روش-های یادگیری ماشین استفاده نمودند. بدین منظور پنج روش مختلف شامل شبکه عصبی مصنوعی (ANN)، سیستم استنتاج فازی عصبی-تطبیقی (ANFIS)، ماشین یادگیری نیرومند (ELM)، رگرسیون فرآیند گاوسی (GPR) و ماشین بردار پشتیبانی (SVM) به کار برده شد و عملکرد آنها براساس چهار شاخص آماری مورد ارزیابی قرار گرفت. نتایج نشان داد که پنج روش فوق می‌توانند پیش‌بینی‌های رضایت بخشی ارائه دهند، در حالی که روش‌های SVM، GPR و ELM از عملکرد بهتری نسبت به ANN و ANFIS در هر دو مرحله آموزش و آزمون برخوردار هستند (Niu and Feng, 2021). فتح آبادی و همکاران کارایی مدل‌های رگرسیون فرآیند گاوس (GPR) و k- نزدیک‌ترین همسایه (k-NN) را در پیش‌بینی جریان رسوبی رودخانه‌های آراز کوسه و اوغان استان گلستان و رودخانه جاجرود استان البرز مورد بررسی قرار دادند. ایشان از داده‌های دبی روز جاری، یک روز قبل و دو روز قبل به همراه مولفه‌های جریان آهسته و سریع به عنوان ورودی استفاده نمودند. نتایج نشان داد که در هر سه رودخانه مورد مطالعه، مدل‌های GPR و k-NN از قدرت پیش‌بینی قابل قبولی برخودار هستند. علاوه بر این، بهترین نتایج برای حوضه‌های ارازکوسه و اوغان زمانی حاصل شد که از داده‌های جریان آهسته و سریع به عنوان ورودی مدل استفاده شدند (Fathabadi et al., 2022). در پژوهشی دیگر ایکرام و همکاران هفت روش یادگیری ماشین را برای پیش‌بینی جریان رودخانه بکار بردند. نتایج به دست آماده حاکی از آن بود که مدل GPR نسبت به سایر روش‌ها از توانایی بهتری در برآورد جریان رودخانه برخوردار بود. البته در برخی از موارد مدل SVR با ورودی‌های یکسان نتایج بهتری را ارائه می‌کرد اما به طور کلی عملکرد دو مدل SVR و GPR نسبت به سایر روش‌ها بهتر بود (Ikram et al., 2022). از روش GPR در زمینه‌های مختلف علم هیدرولوژی نیز به طور گسترده‌ای استفاده شده است که از آن جمله می‌توان به مطالعات کاچتان اغلو و همکاران، البلتاگی و همکراان و حسین زاده و همکاران اشاره نمود (Katipoğlu et al., 2023; Elbeltagi et al., 2023; Hosseinzadeh et al., 2023).

روش جنگل تصادفی یکی از الگوریتم های پرکاربرد برای پیش-بینی سری‌های زمانی هستند (Li et al. 2019). مدل جنگل تصادفی (RF) ترکیبی از درختان طبقه‌بندی و رگرسیون را برای حل مسائل غیرخطی به خدمت گرفته و اغلب عملکرد پیش‌بینی بهتری را در مقایسه با سایر روش‌ها به دست می‌آورد (Li et al. 2019). علاوه بر قدرت پیش‌بینی رضایت‌بخش، مزیت دیگر این روش، پیش‌پردازش ساده داده‌ها است که کاربرد آن را تسهیل می‌کند (Li et al. 2018). جنگل تصادفی می‌تواند الگوهای پیچیده را یاد گرفته و ارتباط غیرخطی بین متغیرهای مستقل و وابسته را در نظر بگیرد. همچنین می‌تواند انواع مختلف داده‌ها را در تجزیه و تحلیل گنجانده و ترکیب کند که این هم به علت عدم وجود توزیع پیش‌فرض درباره داده‌های استفاده شده، می‌باشد. جنگل‌های تصادفی می‌تواند متغیرهای متعدد ورودی را دریافت و موثرترین پارامتر را در مدل‌سازی تشخیص دهد. مزایای مذکور موجب شده است که در مطالعات مختلفی محققان از این روش در پیش‌بینی متغیرهای هیدرولوژیک استفاده کنند (Ahmadi et al. 2022). لی و همکاران با استفاده از روش جنگل-های تصادفی جریان روزانه رودخانه را برآورد نمودند. نتایج نشان داد که مدل RF در پیش‌بینی جریان‌های اوج نسبت به سایر مدل‌ها برتری داشته اما در برآورد مقادیر کمینه مدل ELM بالاترین دقت پیش‌بینی را به خود اختصاص داده بود (Li et al., 2019). علی و همکاران در مطالعه خود از روش‌های رگرسیون ریج و RF برای پیش‌بینی جریان ماهانه رودخانه بهره برده و گزارش نمودند که روش جنگل‌های تصادفی در پیش‌بینی جریان میان مدت با روش‌های پیش‌پردازش داده‌ها عملکرد بهتری دارد (Ali et al., 2020). عبدا و همکاران در پژوهش خود مدل‌های مبتنی بر یادگیری ماشین را برای پیش‌بینی جریان روزانه رودخانه در یک حوضه ساحلی مدیترانه واقع در شمال الجزایر، مورد ارزیابی قرار دادند. بدین منظور از ترکیب‌های ورودی مقادیر بارندگی فعلی و گذشته و مقادیر قبلی جریان روزانه

1 - Gaussian Regression Process (GPR)
2 - Random Forests (RF)

استفاده شد. ایشان از آماره‌های ریشه میانگین مربعات خطا (RMSE) و ضریب همبستگی (R) برای ارزیابی دقت مدل‌ها بهره بردند. تجزیه و تحلیل نتایج نشان داد که RF بهترین نتایج را برای آموزش (RMSE = 4.7 و R = 0.98) و آزمون (RMSE = 2.36 و R = 0.97) ارائه می‌دهد (Abda et al., 2022). لطیف اغلو گزارش نمود که مدل RF برای زمان‌های پیش‌بینی میان مدت و طولانی‌تر جریان رودخانه ازعملکرد بسیار بهتری برخوردار است (Latifoğlu, 2022).

یکی از چالش‌هایی که همچنان در پیش‌بینی جریان رودخانه موثر بوده و عملکرد مدل‌ها را از طریق افزایش عدم قطعیت تحت تاثیر قرار می‌دهد فعالیت‌های انسانی است. این فعالیت‌ها به همراه تاثیراتی که تغییراقلیم بر فرآیندهای هواشناسی می‌گذارد موجب گردیده که داده‌های ثبت شده جریان رودخانه به شدت متغیر و نایستا شوند (Meng et al., 2021). برای حل این مشکل مدل‌های یادگیری ماشین براساس روش‌های تجزیه پایه توسعه داده شده‌اند. این روش‌ها می‌توانند سری‌های جریان رودخانه را به زیرسری‌های متعددی تجزیه نمایند. در این صورت می‌توان خاصیت تناوبی، روند و نویز را در داده-ها شناسایی نمود و اطلاعاتی را که مدل‌ها قادر به استفاده غیرمستقیم از آنها نیستند، استخراج کرد. با وارد کردن زیرسری ها به مدل‌ها عملکرد آنها به طرز چشم‌گیری بهبود می‌یابد (Fang et al., 2019). یکی از روش‌هایی که می‌تواند نا ایستایی داده‌ها را شناسایی کند تبدیل موجک است. این روش با تجزیه سری اصلی به فرکانس‌های بالا و پایین امکان شناسایی تغییرات را برای مدل فراهم می‌کند. یکی از مشکلات اصلی روش تبدیل موجک وابسته بودن آن به تابع موجک مادر می‌باشد به طوری که هر تابع موجک مادر خروجی‌های متفاوتی را برای یک سری ارائه می‌دهد و تشخیص تابع موجک مادر مناسب نیازمند آزمون خطا بوده و زمان بیشتری باید صرف شود (Ahmadi et al., 2022). روش تجزیه مد تجربی (EMD) برخلاف تابع موجک، نیازی به موجک مادر نداشته و توسط هانگ و همکاران (Huang et al., 1998) ارائه شده و تاکنون پیشرفت‌های تکاملی فراوانی را طی کرده است. بعد از معرفی روش اولیه، به دلیل مسائل و مشکلات مربوط به ترکیب (اختلاط) مد، وو و هوانگ روش تجزیه مد تجربی گروهی (EEMD) را معرفی نمودند. روش EEMD دارای مشکلاتی از قبیل باقی ماندن مقداری نوفه در داده بازسازی شده از مدهای ذاتی و نیز تولید مدهای ذاتی مختلف در اثر اضافه کردن نوفه‌های گوسی متفاوت است (Wu and Huang, 2009). از این رو، تورس و همکاران با معرفی تجزیه مد تجربی گروهی کامل (CEEMD) مشکل روش‌های پیشین را برطرف نمودند. در این روش نیز چندین بار نوفه سفید به سیگنال اصلی اضافه می‌گردد، با این تفاوت که در مرحله اول ضمن اضافه کردن نوفه سفید، IMF[1] اول استخراج شده و برای استخراج IMFهای بعدی با اضافه کردن مجدد نوفه مراحل به همین منوال تکرار می گردد (Torres et al., 2011).

با توجه به پیشینه پژوهش ارائه شده، مشاهده می‌شود که برآورد جریان میان‌مدت رودخانه‌ها همواره مورد توجه پژوهش‌گران بوده و بدین منظور از روش‌های یادگیری ماشین و به ویژه RF و GPR استفاده شده است. در اکثر مطالعات انجام شده روش‌های RF و GPR به صورت منفرد مورد استفاده قرار گرفته و ترکیب آنها با روش‌های پیش‌پردازش کننده نظیر CEEMD کمتر مورد توجه بوده است. بنابراین هدف از مطالعه حاضر توسعه مدل‌های CEEMD-RF و CEEMD-GPR برای پیش‌بینی جریان ورودی به سد دز و ارزیابی عملکرد آنها می‌باشد. علاوه بر این در توسعه الگوهای ورودی از روش تئوری آنتروپی شانون برای انتخاب موثرترین داده‌ها بهره گرفته شد.

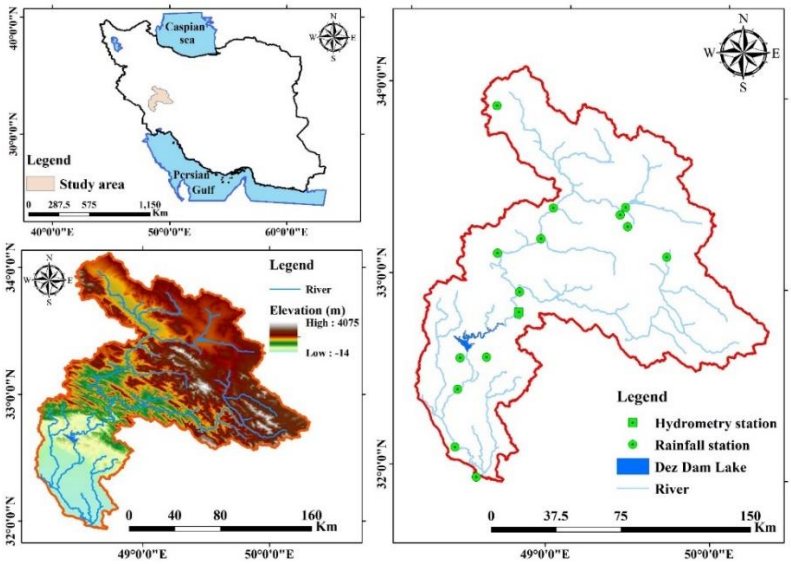
مواد و روش‌ها

منطقه مورد مطالعه و داده‌ها

حوضه آبریز رودخانه دز، یک حوضه درجه سه بوده و در اشل بزرگتر زیرمجموعه‌ای از حوضه‌های آبریز کارون بزرگ و خلیج فارس و دریای عمان می‌باشد. مساحت کل حوضه بالغ بر ۲۱۷۲۰ کیلومتر مربع و متوسط ارتفاع حوضه حدود ۱۶۰۰ متر می‌باشد. حوضه آبریز دز چهار زیرحوضه تیره، ماربره، سزار و بختیاری را شامل می‌شود. دو شاخه سزار و بختیاری در محلی به نام تنگ پنج به یکدیگر رسیده و رودخانه دز را تشکیل می‌دهند. بر روی رودخانه دز سدی به همین نام احداث شده است. سد دز یکی از سدهای قدیمی و بزرگ ایران و یک سد بتنی برق‌آبی از گونه دو قوسی جدار نازک می‌باشد. این سد بزرگ در سال ۱۳۳۸ خورشیدی در ۲۳ کیلومتری شهر اندیمشک استان خوزستان در کوه‌های زاگرس آغاز و در سال ۱۳۴۱ خورشیدی پایان یافت. سد دز ۱۲۵۰۰۰ هکتار از ارضی پایین‌دست را آبیاری می‌کند و نقش به سزایی در مدیریت سیلاب‌های بالادست دارد. نیروگاه این سد بزرگ دارای قدرت ۵۲۰ مگاوات می‌باشد. دریاچه سد طولی برابر با ۶۵ کیلومتر داشته و مساحت آن حدود ۶۰ کیلومترمربع بوده و گنجایش آن ۳/۳ میلیارد مترمکعب می‌باشد. هدف اصلی از احداث سد دز تامین نیروی لازم جهت تولید برق، کنترل سیلاب‌ها و تنظیم آب جهت مصارف آبیاری است. جریان ورودی به سد دز توسط ایستگاه هیدرومتری تله زنگ اندازه‌گیری می‌شود. این ایستگاه از سال ۱۳۳۴ فعال بوده و مساحتی بالغ بر ۱۶۲۰۱ کیلومتر مربع را تحت پوشش خود قرار می‌دهد. در بالادست سد دز نیز ۱۱ ایستگاه باران‌سنجی وجود دارد که در خلال سال‌های ۱۳۳۴ تا ۱۳۵۰ تاسیس شده‌اند. با بررسی‌های انجام شده دوره آماری برای مدل‌سازی جریان ورودی به سد دز بین سال‌های آبی ۱۳۵۰ تا ۱۳۹۶ در نظر گرفته شد.

1 - Intrinsic Mode Function

مشخصات آماری ایستگاه‌های مورد مطالعه در جدول ۱ ذکر گردیده است. شکل ۱ نیز موقعیت حوضه آبریز دز در ایران و پراکنش ایستگاه‌ها در سطح حوضه را نشان می‌دهد.



شکل ۱- موقعیت حوضه آبریز دز در ایران و پراکنش ایستگاه‌های مورد مطالعه

جدول ۱- مشخصات جغرافیایی و آماری ایستگاه های مورد بررسی در دوره آماری ۱۳۵۰ تا ۱۳۹۶

ردیف	نوع ایستگاه	نام ایستگاه	ارتفاع (متر)	UTM (X)	UTM (Y)	میانگین سالانه	یکا
۱	هیدرومتری	تله زنگ	۴۶۸	۲۹۱۰۲۱	۳۶۳۳۷۶۳	۲۴۷/۹	m^3/s
۲		سد دز	۵۲۵	۲۵۷۸۷۵	۳۶۰۶۹۲۰	۴۶۷/۸	
۳		ونایی	۲۰۰۰	۲۷۵۱۲۷	۳۷۵۳۷۳۱	۷۱۶/۶	
۴		چم زمان	۱۸۳۰	۳۵۱۲۰۵	۳۶۹۶۷۷۶	۴۹۰/۵	
۵		کمندان	۱۹۳۰	۳۵۲۷۱۶	۳۶۸۵۶۶۲	۷۰۸/۸	
۶		دره تخت	۱۸۹۰	۳۴۸۱۶۴	۳۶۹۲۳۸۷	۷۲۶/۷	
۷	باران سنجی	چم چیت	۱۲۹۰	۳۰۹۴۱۰	۳۶۹۵۲۹۰	۷۱۸/۴	mm/year
۸		سپید دشت- سزار	۹۷۰	۳۰۲۷۳۶	۳۶۷۷۳۰۴	۷۱۴/۳	
۹		کشور	۷۷۰	۲۷۷۷۹۰	۳۶۶۸۲۲۴	۹۵۰/۷	
۱۰		تنگ پنج- بختیاری	۵۴۰	۲۹۱۱۹۲	۳۶۴۶۱۰۹	۱۱۵۱/۴	
۱۱		تله زنگ	۴۴۰	۲۹۰۹۱۹	۳۶۳۳۱۷۲	۸۶۴/۹	
۱۲		کاظم آباد	۲۰۰۰	۳۷۵۷۸۲	۳۶۶۸۷۰۶	۶۰۷/۹	

روش جنگل‌های تصادفی (RF)

جنگل تصادفی (RF) یک روش یادگیری ماشینی است که توسط بریمن در سال ۲۰۰۱ توسعه یافته است (Breiman, 2001). به دلیل کارکرد مناسب و تعمیم آسان، این روش به طور گسترده در زمینه‌های مختلف استفاده شده است (Were et al. 2015). RF یک روش مجموعه‌ای است، به این معنی که در ابتدا چندین درخت (یک جنگل) ساخته می‌شود و در گام بعد مقادیر پیش بینی شده با ترکیبی از نتایج همه مدل‌های درختی در جنگل به دست می‌آید. الگوریتم RF برای رگرسیون n مجموعه داده از نمونه‌های تصادفی را با جایگزینی (بوت استرپ) از مجموعه داده اصلی استخراج می‌کند (Orellana-Alvear et al., 2020). سپس، از این مجموعه داده‌های جدید برای ساختن n درخت استفاده کرده بدین طریق تضمین می‌کند که زیر مجموعه متفاوتی برای ساخت هر درخت تصمیم در مدل استفاده شود. به عبارت دیگر RF ترکیبی از درخت‌های تصمیم ساده است که در آن هر درخت با انتخاب نمونه‌ها و ویژگی‌های تصادفی از همه پیش‌بینی‌کننده‌ها تولید می‌شود (Orellana-Alvear et al., 2020).

به منظور ایجاد درخت رگرسیونی از جزءبندی بازگشتی و رگرسیون-های چندگانه استفاده می‌شود. فرایند تصمیم در هر گره داخلی از گره ریشه، طبق قانون درختی تکرار می‌شود تا زمانی که شرط توقف تعیین شده بدست آید. (Breiman, 2001). در روش RF بردار تصادفی X_n که مستقل از بردارهای تصادفی $X_1, X_2, ..., X_{n-1}$ بوده، برای درخت nام تولید می‌شود. همچنین همه بردارها از توزیع مشابهی تبعیت می‌کنند. رگرسیون درختی با استفاده از مجموعه داده‌های آموزش و X_n محاسبه شده مجموعه درختهایی برابر با n را به شرح زیر تولید می‌نماید (Breiman, 2001):

$$X_n = \{h_1(x), h_2(x), ..., h_n(x)\} \quad (۱)$$

$$h_n = h(x, X_n), x = \{x_1, x_2, ..., x_p\} \quad (۲)$$

بردار p بعدی فوق یک جنگل را تشکیل داده و خروجی‌ها برای هر درخت به صورت زیر ارائه می‌شود:

$$y_1 = h_1(x), y_2 = h_2(x), ..., y_n = h_n(x) \quad (۳)$$

که در رابطه فوق، y_n خروجی درخت nام می‌باشد. برای به دست آوردن خروجی نهایی، متوسط همه پیش‌بینی‌های درخت‌ها محاسبه می‌شود (Breiman, 2001). خطای پیشبینی نیز طبق رابطه ۴ محاسبه می‌گردد.

$$MSE = \frac{\sum_{i=1}^{n}\left[y(x_i) - y_i \right]^2}{n} \quad (۴)$$

در رابطه فوق $y(x_i)$ نتایج محاسباتی، y_i نتایج مشاهداتی و n تعداد کل مشاهدات است و MSE میزان خطای بین مقادیر مشاهداتی و محاسباتی را نشان می‌دهد. اکثر نتایج جمع‌آوری شده از درختان تصمیم به عنوان خروجی نهایی RF در نظر گرفته می‌شود (Al-Abadi and Shahid 2016). به طور کلی، بخش جداگانه‌ای از نمونه‌ها باید از مجموعه داده اصلی حذف شود که به آنها نمونه‌های خارج از کیسه (OOB) اطلاق می‌گردد. میانگین مربعات خطای نمونه‌های OOB (رابطه ۴) اغلب برای ارزیابی عملکرد روش RF استفاده می‌شود. دو پارامتر مهم وجود دارد که باید در طول فرآیند واسنجی مدل RF تعیین شوند: الف) تعداد درختان موجود در جنگل (n_{tree}) و ب) تعداد پیش‌بینی‌کننده‌های آزمایش‌شده در هر گره (m_{try}). برای به دست آوردن حداقل خطای تعمیم و همبستگی بین درخت-های تصمیم، مقدار m_{try} بر اساس روابط تجربی زیر قابل محاسبه می‌باشد (Huang et al., 2016; Al-Abadi and Shahid, 2016):

$$\begin{cases} m_{try} = \log_2(M+1) \\ m_{try} = \sqrt{M} \\ m_{try} = \dfrac{M}{3} \end{cases} \quad (۵)$$

در روابط فوق M تعداد متغیرهای ورودی تعریف شده در مجموعه داده اصلی است. ارزش n_{tree} به طور قابل توجهی بر نتایج پیش بینی تاثیر می‌گذارد. بنابراین، آزمایشی برای انتخاب معقول‌ترین درخت طراحی می‌شود. با شروع از مقدار اولیه n_{tree} = 1، مدل RF با تعداد فزاینده‌ای از درختان آموزش داده شده و آزمایش می‌شود تا زمانی که n_{tree} = 500 (با اندازه مرحله ۱) به دست آید. در این مطالعه از نرم افزار متن باز WEKA3.9 برای توسعه و طراحی مدل RF استفاده شد.

رگرسیون فرآیند گاوسی (GPR)

مدل GPR یک مدل ناپارامتری احتمالاتی مبتنی بر کرنل بوده و بر این فرض استوار است که در GPR هر نمونه یادگیری از احتمالات قبلی فرآیند گاوسی پیروی کرده و بر این اساس می‌توان احتمال پسین مربوطه را محاسبه کرد. GPR از تابع کرنل برای تعریف کوواریانس توزیع قبلی بر روی توابع هدف استفاده می‌کند. بدین منظور یک تابع درستنمایی با توجه به داده‌های آموزشی مشاهده شده تعریف می‌گردد. به عبارت دیگر، یک توزیع پسین (گاوسی) بر روی توابع هدف تعریف شده است که میانگین آن برای پیش‌بینی می‌تواند به کار گرفته شود (Zhu et al., 2019). رگرسیون فرآیند گاوسی مجموعه‌ای است از متغیرهای تصادفی که هر یک دارای توزیع گاوسی مشترک می‌باشند. تحلیل رگرسیون GPR این امکان را فراهم می‌نماید تا راهی برای طبقه‌بندی داده‌ها براساس ساختارهایی که در آنها وجود دارد، صورت پذیرد. در فرآیند GPR تابع f به عنوان یک تابع توزیع تعریف می‌گردد. در این تابع، f نگاشتی از فضای ورودی Y به فضای R است و برای هر زیر مجموعه متناهی از Y، توزیع حاشیه ای آن به صورت $p(f(y1),f(y_2),\ldots, f(y_n))$ قابل تعریف بوده که یک تابع توزیع نرمال چندمتغیره می‌باشد (Zhu et al., 2019). فرآیند گاوسی پارامتریک با استفاده از میانگین $y(m)$ و کوواریانس $k(y_i,y_j)$ به شرح زیر محاسبه می‌شود:

$$f\left|Y \square N(m(y), K(Y_i, Y_j)\right| \quad (۶)$$

شکل ریاضی رابطه فوق به صورت زیر قابل بازنویسی است (Zhu et al., 2019):

$$f(x) \sim gp\left(m(y),\ k\left(y_i\ ,\ y_j\right)\right) \quad (۷)$$

در این رابطه ردیف‌ها در ماتریس y همان بردارهای ورودی بوده،

f برداری است از مقادیر تابع و $k\left(y_i\ ,\ y_j\right)$ ماتریس کواریانس $n\times n$ را نشان می‌دهد. حال می‌توان برازش اولیه مدل رگرسیون فرآیند گاوسی را با درنظر گرفتن y به عنوان مشاهده همراه با نوفه ε به شرح زیر بیان نمود:

$$y=f(x)+\varepsilon,\qquad \varepsilon\ \square\ N(0,\sigma_n^2) \tag{۸}$$

در این صورت توزیع مشترک خروجی‌های آموزشی y و داده-های آزمون f^* با تابع متوسط صفر برابر است با:

$$\begin{bmatrix} y \\ f^* \end{bmatrix}\ \square\ N\left(0,\begin{bmatrix} K(Y,Y)+\sigma_n^2 I & K(Y,Y^*) \\ K(Y^*,Y) & K(Y^*,Y^*) \end{bmatrix}\right) \tag{۹}$$

در رابطه فوق Y ماتریس داده‌های آموزشی و Y^* ماتریس داده-های آزمون می‌باشد. با مقید کردن f^* بر روی مشاهده y می‌توان پیش‌بینی توزیع را از طریق رابطه زیر نشان داد:

$$f^*\left|Y,y,Y^*\ \square\ N(\overline{f^*},V(f^*))\right. \tag{۱۰}$$

در این رابطه $\overline{f^*}$ و $V(f^*)$ از روابط ۱۱ و ۱۲ قابل محاسبه می‌باشند:

$$\overline{f^*}=K(Y^*,Y)\left[\ K(Y,Y)+\sigma_n^2 I\ \right]^{-1} y \tag{۱۱}$$

$$V(f^*)=K(Y^*,Y^*)-K(Y^*,Y)\left[\ K(Y,Y)+\sigma_n^2 I\ \right]^{-1} K(Y,Y^*) \tag{۱۲}$$

معادله ۱۱ نشان دهنده آن است که میانگین مقادیر تخمینی تابعی خطی از مشاهدات نویزدار y می‌باشد. به طور کلی GPR با این فرض توسعه داده شده است که احتمالا ورودی‌های نزدیک به هم، خروجی‌های مشابهی را نتیجه خواهند داد، بنابراین به نمونه‌های دارای مقادیر یکسان وزن بیشتری اختصاص خواهد یافت.

تئوری آنتروپی شانون

همانطور که در نظریه اطلاعات تعریف شده است، آنتروپی اندازه گیری مقدار اطلاعات مورد نیاز برای توصیف یک متغیر تصادفی است. به عبارت دیگر، آنتروپی نشان دهنده مقدار عدم قطعیت است که توسط توزیع احتمال یک متغیر تصادفی ارائه می‌شود. اساس آنتروپی شانون این است که اطلاعات به دست آمده از یک رویداد با احتمال وقوع p برابر است با $log\ (1/p)$. از این رو می‌توان نتیجه گرفت که عدم قطعیت پیش‌بینی یک رویداد با احتمال آن رابطه عکس دارد (Darbandsari and Coulibaly, 2020). تابع لگاریتمی، به عنوان یک تابع انتقالی عمل کرده و این اطمینان را ایجاد می‌کند که اطلاعات به دست آمده از وقوع مشترک دو رویداد مستقل برابر است با مقدار اطلاعاتی که هر رخداد به تنهایی ارائه می‌دهد. این رابطه به شرح زیر بیان می‌شود:

$$\log\left(\frac{1}{p(x_1,x_2)}\right)=\log\left(\frac{1}{p(x_1)}\right)+\left(\frac{1}{p(x_2)}\right)=-\log(p(x_1))-\log(p(x_2)) \tag{۱۳}$$

در نتیجه می‌توان برای متغیرهای گسسته (X) با k داده و احتمالات ($p_1, p_2,..., p_k$) متوسط اطلاعات X از رابطه زیر به دست می‌آید:

$$H(X)=E\left(\log\left(\frac{1}{p_1,p_2,...,p_k}\right)\right)=-\sum_{i=1}^{k}p_i\log(p_i) \tag{۱۴}$$

در رابطه فوق $E(.)$ تابع انتظار و H(X) آنتروپی حاشیه‌ای متغیر X در سیستم دودویی است، زیرا مبنای لگاریتم برابر با عدد دو در نظر گرفته می شود. در یک ماتریس تصمیم گیری Pi می‌تواند برای ارزیابی گزینه‌های مختلف به کار رود. در ماتریس تصمیم‌گیری زیر m گزینه و n شاخص (معیار) مدنظر می‌باشند (Saray et al., 2020).

$$D=\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \tag{۱۵}$$

آنتروپی Ej به شرح زیر محاسبه می‌گردد (Saray et al., 2020):

$$E_j=-K\sum_{i=1}^{m}P_{ij}LnP_{ij}\ ;\ \forall_{ij} \tag{۱۶}$$

و K به عنوان مقدار ثابت به صورت زیر محاسبه می‌گردد:

$$K=\frac{1}{Lnm} \tag{۱۷}$$

در ادامه مقدار dj (درجه انحراف) محاسبه می‌شود که بیان می-کند شاخص مربوطه (j) چه میزان اطلاعات مفید برای تصمیم‌گیری در اختیار تصمیم‌گیرنده قرار می‌دهد (Shannon, 2001).

$$d_j=1-E_j;\ \forall_j \tag{۱۸}$$

در نهایت مقدار وزن Wj محاسبه می‌گردد که در آن بزرگترین وزن نشان دهنده اهمیت پارامتر مورد نظر می‌باشد (Shannon, 2001):

$$W_j=\frac{d_j}{\sum_{i=1}^{n}d_j};\ \forall_j \tag{۱۹}$$

روش تجزیه مد تجربی یکپارچه کامل (CEEMD)

هوانگ و همکاران روش تجزیه مود تجربی (EMD) را که یک روش تحلیلی و کارآمد غیر خطی و غیر ثابت در تجزیه فرکانس داده-های زمانی است در سال ۱۹۹۸ معرفی نمودند (Huang et al., 1998). اساس این روش براین فرض استوار است که هر سیگنال دارای زیرمجموعه های مختلفی تحت عنوان توابع مد ذاتی (IMF)

می‌باشد (Keshvari et al., 2022). توابع مد ذاتی (IMF) باید بتوانند دو شرط زیر را تامین نمایند: الف) در کل داده‌ها، تعداد نقاط اکسترمم و نقاط صفر بایستی با هم برابر و یا حداکثر دارای یک واحد اختلاف باشند و ب) در هر نقطه میانگین پوش برازش داده شده بر نقاط بیشینه محلی و پوش برازش داده شده بر نقاط کمینه محلی باید برابر با صفر باشد.

در بعضی از موارد بدلیل اختلاف در مدها، سیگنال‌های دچار تناوب و نویز شده و باعث ایجاد گسیختگی توزیع حوزه زمان–فرکانس می‌گردد. در نتیجه این امر میانگین IMFها مبهم شده و EMD نمی‌تواند بدرستی عمل نماید. به منظور حل این مشکل، تجزیه مد تجربی گروهی کامل (CEEMD) توسط ارائه گردید و دارای مراحل زیر است (Torres et al., 2011):

در گام اول، در هر مرحله درصدی از نویز گوسین به داده‌ها اضافه می‌شود. برای I نویز مختلف گوسی EMD انجام شده و مقدار IMF اول از میانگین I مد به دست می‌آید.

$$IMF_1 = \frac{1}{I}\sum_{i=1}^{I} E_1\left[x + \varepsilon w_i\right] \qquad (۲۰)$$

سپس مقدار r_1 از رابطهٔ زیر محاسبه می‌شود:

$$r_1 = x - IMF_1 \qquad (۲۱)$$

این روند برای محاسبهٔ IMF_{K+1}ام همانند مرحله اول تا زمانی که بیش از دو اکسترمم باقیمانده نداشته باشد تکرار می‌گردد. پس از محاسبه IMF ها، به کمک تبدیل هیلبرت، فرکانس لحظه ای بدست آمده و به شکل تصویری دوبعدی از فرکانس لحظه ای متغیر نمایش داده می شود (Keshvari et al., 2022)

توسعه الگوی‌های ورودی برای مدل‌ها

یکی از گام‌های مهم در مدل‌سازی پدیده‌های پیچیده هیدرولوژیک، تعیین و انتخاب اطلاعات تاثیرگذار در پدیده مورد بررسی می‌باشد. این اطلاعات برای آموزش ماهیت ساز و کار حاکم به مدل‌ها بسیار حائز اهمیت بوده و عملکرد مدل را تحت تاثیر قرار می‌دهد. اما ارائه اطلاعات بیش از اندازه به مدل هیچ تضمینی را برای دستیابی به نتایج دقیق‌تر ارائه نمی‌دهد. دلیل این امر آن است که افزایش تعداد ورودی‌ها باعث پیچیدگی الگوها و افزایش حافظه درگیر شده و در نهایت کاهش دقت را به همراه خواهد داشت. بنابراین در پیش‌بینی جریان ورودی به سد دز بایستی سعی نمود از موثرترین داده‌های مشاهداتی برای آموزش مدل‌های RF و GPR بهره برد. در مطالعه حاضر سعی گردید الگوهای ورودی ابتدا مورد پردازش قرار گرفته و در گام بعد به مدل‌های یادگیری ماشین معرفی گردد. در ساختار طراحی شده، ابتدا الگوهای ورودی براساس سناریوهای مشخص به مدل‌های RF و GPR ارائه می‌شود. در گام بعد بهترین الگوها مشخص شده و با روش CEEMD پردازش گردیده و مدل‌های تلفیقی CEEMD-RF و CEEMD-GPR توسعه داده می‌شوند.

در این مطالعه، برای پیش‌بینی جریان ورودی ماه فعلی به سد دز، از الگوهایی استفاده شد که براساس داده‌های جریان ماه‌های قبل، بارش ماه‌های پیشین و دخالت دادن خاصیت تناوبی توسعه یافته‌اند. در بالا دست سد دز ۱۱ ایستگاه باران‌سنجی اطلاعات بارش را ثبت می‌کنند که در این صورت امکان استفاده از همه آنها در فرآیند مدل سازی وجود ندارد. بنابراین در پژوهش حاضر از روش تئوری آنتروپی شانون برای انتخاب ایستگاه باران‌سنج مناسب برای پیش‌بینی جریان ورودی به سد استفاده شد. در گام بعد نمودار خودهمبستگی جزئی سری‌های زمانی جریان ماهانه ورودی به سد دز در محل ایستگاه هیدرومتری تله زنگ رسم و مورد بررسی قرار گرفت (شکل ۲). با توجه به شکل ۲ مشاهده می‌شود که تا شش تاخیر داده‌های جریان به یکدیگر وابستگی معنی‌دار دارند و از این‌رو الگوهای ورودی تا شش تاخیر آماده شدند. همچنین شکل ۲ نشان می‌دهد که خاصیت تناوبی در جریان ماهانه ثبت شده در ایستگاه تله زنگ وجود داشته و بنابه پیشنهاد منتصری و قویدل از شماره ماه‌ها برای معرفی اثر پریودیک استفاده شد (منتصری و قویدل، ۱۳۹۳). به طور کلی ساختار الگوهای ورودی در چهار حالت تاخیر ماهانه دبی، تاخیر ماهانه دبی و بارش و در نظر گرفتن ترم تناوبی برای دوسناریوی قبل تعریف و در جدول ۲ ارائه شد.

ارزیابی مدل‌ها

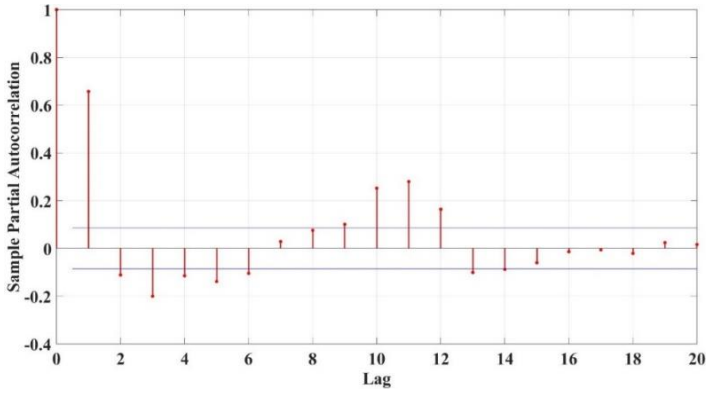
در این پژوهش، برای ارزیابی مدل‌های به کار گرفته شده در پیش بینی جریان ورودی به سد دز از معیارهای میانگین جذر مربعات خطا (RMSE)، میانگین خطای مطلق (MAE)، کلینگ گوپتا (KGE) و ضریب ویلموت (WI) استفاده می‌شود:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(O_i - P_i\right)^2}{n}} \qquad (۲۲)$$

$$MAE = \left|\frac{O_i - P_i}{n}\right| \qquad (۲۳)$$

$$KGE = 1 - \sqrt{\left(CC-1\right)^2 + \left(\alpha-1\right)^2 + \left(\beta-1\right)^2} \qquad (۲۴)$$

$$WI = \left|1 - \left[\frac{\sum_{i=1}^{n}\left(O_i - P_i\right)^2}{\sum_{i=1}^{n}\left(\left|P_i - \bar{O}\right| + \left|O_i - \bar{O}\right|\right)^2}\right]\right|, 0 \le WI \le 1 \qquad (۲۵)$$



شکل ۲- نمودار خودهمبستگی جزئی جریان ماهانه وردی به سد دز در محل ایستگاه تله زنگ

جدول ۲- الگوهای ورودی به مدل های منفرد RF و GPR

سناریو	ردیف	الگو	آرایش الگوی ورودی
تاخیر داده‌های جریان ورودی به سد (S1)	۱	S1M1	$Q_t = f(Q_{t-1})$
	۲	S1M2	$Q_t = f(Q_{t-1}, Q_{t-2})$
	۳	S1M3	$Q_t = f(Q_{t-1}, Q_{t-2}, Q_{t-3})$
	۴	S1M4	$Q_t = f(Q_{t-1}, Q_{t-2}, Q_{t-3}, Q_{t-4})$
	۵	S1M5	$Q_t = f(Q_{t-1}, Q_{t-2}, Q_{t-3}, Q_{t-4}, Q_{t-5})$
	۶	S1M6	$Q_t = f(Q_{t-1}, Q_{t-2}, Q_{t-3}, Q_{t-4}, Q_{t-5}, Q_{t-6})$
تاخیر داده‌های بارش و جریان ورودی به سد (S2)	۷	S2M1	$Q_t = f(R_{t-1}, Q_{t-1})$
	۸	S2M2	$Q_t = f(R_{t-1}, QR_{t-2}, Q_{t-1}, Q_{t-2})$
	۹	S2M3	$Q_t = f(R_{t-1}, R_{t-2}, R_{t-3}, Q_{t-1}, Q_{t-2}, Q_{t-3})$
	۱۰	S2M4	$Q_t = f(R_{t-1}, R_{t-2}, R_{t-3}, R_{t-4}, Q_{t-1}, Q_{t-2}, Q_{t-3}, Q_{t-4})$
	۱۱	S2M5	$Q_t = f(R_{t-1}, R_{t-2}, R_{t-3}, R_{t-4}, R_{t-5}, Q_{t-1}, Q_{t-2}, Q_{t-3}, Q_{t-4}, Q_{t-5})$
	۱۲	S2M6	$Q_t = f(R_{t-1}, R_{t-2}, R_{t-3}, R_{t-4}, R_{t-5}, R_{t-6}, Q_{t-1}, Q_{t-2}, Q_{t-3}, Q_{t-4}, Q_{t-5}, Q_{t-6})$
تاخیر داده‌های جریان ورودی به سد با در نظر گرفتن ترم تناوبی (S3)	۱۳	S3M1	$Q_t = f(\propto, Q_{t-1})$
	۱۴	S3M2	$Q_t = f(\propto, Q_{t-1}, Q_{t-2})$
	۱۵	S3M3	$Q_t = f(\propto, Q_{t-1}, Q_{t-2}, Q_{t-3})$
	۱۶	S3M4	$Q_t = f(\propto, Q_{t-1}, Q_{t-2}, Q_{t-3}, Q_{t-4})$
	۱۷	S3M5	$Q_t = f(\propto, Q_{t-1}, Q_{t-2}, Q_{t-3}, Q_{t-4}, Q_{t-5})$
	۱۸	S3M6	$Q_t = f(\propto, Q_{t-1}, Q_{t-2}, Q_{t-3}, Q_{t-4}, Q_{t-5}, Q_{t-6})$
تاخیر داده‌های بارش و جریان ورودی به سد با در نظر گرفتن ترم تناوبی (S4)	۱۹	S4M1	$Q_t = f(\propto, R_{t-1}, Q_{t-1})$
	۲۰	S4M2	$Q_t = f(\propto, R_{t-1}, QR_{t-2}, Q_{t-1}, Q_{t-2})$
	۲۱	S4M3	$Q_t = f(\propto, R_{t-1}, R_{t-2}, R_{t-3}, Q_{t-1}, Q_{t-2}, Q_{t-3})$
	۲۲	S4M4	$Q_t = f(\propto, R_{t-1}, R_{t-2}, R_{t-3}, R_{t-4}, Q_{t-1}, Q_{t-2}, Q_{t-3}, Q_{t-4})$
	۲۳	S4M5	$Q_t = f(\propto, R_{t-1}, R_{t-2}, R_{t-3}, R_{t-4}, R_{t-5}, Q_{t-1}, Q_{t-2}, Q_{t-3}, Q_{t-4}, Q_{t-5})$
	۲۴	S4M6	$Q_t = f(, R_{t-1}, R_{t-2}, R_{t-3}, R_{t-4}, R_{t-5}, R_{t-6}, Q_{t-1}, Q_{t-2}, Q_{t-3}, Q_{t-4}, Q_{t-5}, Q_{t-6})$

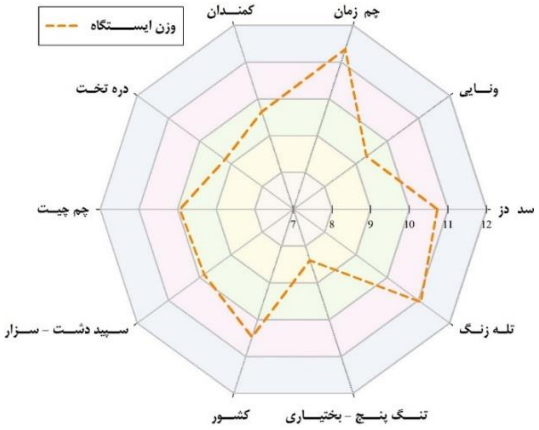
که در روابط فوق، O_i مقادیر مشاهداتی، P_i مقادیر پیش‌بینی شده، $\bar{O}$ میانگین جریان مشاهداتی، CC ضریب همبستگی بین داده-های مشاهداتی و محاسباتی، α نسبت انحرف معیار O_i و P_i، β نسبت میانگین O_i و P_i و n تعداد داده‌ها می‌باشد. مدلی به عنوان مناسب‌ترین گزینه انتخاب می‌گردد که کمترین (بیشترین) مقدار RMSE و MAE (KGE و WI) را به خود اختصاص دهد (Willmott et al., 2012).

نتایج

انتخاب ایستگاه باران‌سنجی مناسب برای پیش‌بینی جریان ورودی به سد دز

همانگونه که در قسمت مواد و روش‌ها ذکر گردید، در حوضه آبریز دز ۱۱ ایستگاه باران‌سنجی وجود داشته و امکان استفاده از همه آنها در فرآیند مدل‌سازی میسر نمی‌باشد. همچنین به دلیل توپوگرافی به شدت متغیر حوضه آبریز سد دز، کاربرد روش تیسن برای منطقه‌ای کردن بارش‌ها نیز مقدور نیست. بنابراین در این پژوهش وزن هر ایستگاه با استفاده از روش تئوری آنتروپی شانون محاسبه گردید. شکل ۳ وزن هریک از ایستگاه‌ها را نشان می‌دهد. با توجه به این شکل مشاهده می‌شود که ایستگاه چم زمان در مقایسه با سایر ایستگاه‌ها وزن بیشتری را به خود اختصاص داده و از این رو برای ادامه محاسبات انتخاب گردید.

ناظری و همکاران برای پیش‌بینی جریان رودخانه، از تئوری آنتروپی برای انتخاب ایستگاه باران‌سنج موثر در مدل‌سازی جریان رودخانه دز بهره بردند. بر اساس نتایج به دست آمده از آن تحقیق، روش آنتروپی ایستگاه ونایی را به عنوان ایستگاه شاخص معرفی نموده است (Nazer et al., 2022). با بررسی های انجام شده مشاهده گردید که دوره آماری مورد مطالعه در پژوهش فوق، ۳۰ ساله بوده اما در این مطالعه، دوره آماری ایستگاه‌های مورد بررسی ۴۷ ساله می‌باشد. بنابراین تفاوت در طول دوره آماری و ماهیت روش‌های مورد بررسی را می‌توان از دلایل مهم انتخاب دو ایستگاه متفاوت به عنوان ایستگاه شاخص، برای یک حوضه آبریز معرفی نمود.



شکل ۳- نتایج روش تئوری آنتروپی شانون در وزن دهی ایستگاه‌های باران‌سنجی حوضه آبریز سد دز

پیش‌بینی جریان ورودی به سد دز با استفاده از مدل‌های منفرد

در این مطاله مدل‌های مستقل RF و GPR برای مدل‌سازی به کار گرفته شد. بدین منظور از الگوهای ارائه شده در جدول ۲ تحت چهار سناریوی مختلف استفاده و ۸۰ درصد داده‌ها (۵۰۰ ماه) برای آموزش مدل‌ها و ۲۰ درصد (۱۲۰ ماه) برای آزمون مورد استفاده قرار گرفت. در این مطالعه برای اجرای مدل‌های RF و GPR از نرم افزار متن باز WEKA استفاده شد. یکی از گام‌های مهم در کاربرد مدل-های یادگیری ماشین، تعیین پارامترهای تاثیرگذار هر مدل می‌باشد. در روش GPR انتخاب تابع کرنل مناسب و تنظیم پارامترهای کرنل منتخب از مهمترین مراحل مدل‌سازی به شمار می‌رود. در این پژوهش، توابع کرنل چند جمله‌ای (POL) و پایه شعاعی (RBF) برای پیش‌بینی جریان ورودی به سد مورد استفاده قرار گرفتند. رضازاده جودی و ستاری ضمن بررسی توابع کرنل مختلف برای روش GPR، تابع کرنل RBF را برای مسائل پیچیده توصیه نموده‌اند (رضازاده جودی و ستاری، ۱۳۹۵). در این مطالعه پارامترهای مربوط به هر کرنل با استفاده از فرآیند سعی و خطا تعیین گردید. شایان ذکر است که در کلیه الگوهای مورد استفاده کرنل RBF از عملکرد بهتری نسبت به کرنل POL برخوردار بود.

در جداول ۳ و ۴ نتایج تحلیل‌های آماری شاخص‌های ارزیابی RMSE، MAE، KGE و WI به ترتیب برای مدل‌های RF و GPR ارائه شده است. با توجه به این جداول مشاهده می‌شود که سناریوهای تعریف شده در دقت و عملکرد مدل‌های منفرد تاثیر مستقیمی داشته‌اند. به عنوان مثال، برای مدل RF کاربرد ترم تناوبی کارایی مدل را براساس شاخص KGE بهبود بخشیده است به طوری که این شاخص بدون در نظر گرفتن ترم تناوبی برای الگوهای ورودی S1M1 تا S2M6 به طور متوسط برابر با ۰/۴۷ بوده و این در حالی است که برای الگوهای S3M1 تا S4M6 این مقدار به ۰/۵۱ ارتقا یافته است. بهبود در شاخص KGE برای روش GPR نیز با در نظر گرفتن خاصیت تناوبی وجود داشته و حدود ۰/۰۳ بهبود مشاهده می-شود. احمدی در پژوهش خود نشان داد که کاربرد ترم تناوبی می‌تواند

عملکرد مدل‌ها را در پیش‌بینی جریان رودخانه بهبود بخشد (احمدی، ۱۳۹۹). همچنین مطالعات دیگر نظیر منتصری و قویدل کاربرد ترم تناوبی در مدل‌سازی را توصیه نموده‌اند که با نتایج این پژوهش همخوانی دارد (منتصری و قویدل، ۱۳۹۳).

مقایسه شاخص RMSE برای مدل‌های منفرد RF و GPR نشان می‌دهد که مدل فرآیند رگرسیون گوسی در مقایسه با روش جنگل‌های تصادفی از عملکرد بهتری برخوردار بوده و به ورودی‌های کمتری نیز نیاز دارد به طوریکه GPR با الگوی S1M1 کمترین خطا (برابر با ۹۷/۴۹ متر مکعب بر ثانیه) را نتیجه داده اما مدل RF با الگوی S4M2 خطایی برابر با ۱۱۱/۰۰ متر مکعب بر ثانیه را ثبت کرده است. به عبارت دیگر، عملکرد مدل GPR در مقایسه با RF هم از نظر خطا و هم از نظر تعداد داده‌های مورد نیاز برای انجام مدل-سازی جریان ورودی به سد دز بهتر بوده است.

نکته مهم دیگری که با بررسی شاخص‌های ارزیابی مشاهده می-شود تاثیر اندک استفاده از بارش در عملکرد مدل‌های منفرد می‌باشد. در سناریوی دوم که الگوهای ورودی S2M1 تا S2M6 را شامل می‌شود، شاخص MAE مدل RF نسبت به سناریوی اول (با الگوهای ورودی S1M1 تا S1M6) که فقط از حافظه سری جریان ورودی به سد برای پیش‌بینی بهره می‌برد، در بهترین حالت از ۷۵/۱۴ (الگوی S1M1) متر مکعب در ثانیه به ۷۸/۸۹ (الگوی S2M2) افزایش داده است. این روند در عملکرد مدل GPR نیز مشهود است. دلیل این امر را می‌توان در افزایش تعداد لایه‌های ورودی و عدم توانایی مدل‌های منفرد در یافتن ارتباط بین پدیده‌ها جستجو کرد و از این رو استفاده از روش‌های مبتنی بر پیش‌پردازش داده‌ها بسیار می‌تواند مفید واقع شود (Ahmadi et al., 2022).

جدول ۳- نتایج شاخص‌های ارزیابی عملکرد مدل RF در مرحله آموزش و آزمون

مرحله آموزش					مرحله آزمون				
الگو	RMSE (m^3/s)	MAE (m^3/s)	KGE	WI	الگو	RMSE (m^3/s)	MAE (m^3/s)	KGE	WI
S1M1	۱۰۸/۷۷	۶۹/۴۸	۰/۷۶	۰/۸۱	**S1M1**	**۱۱۷/۸۵**	**۷۵/۱٤**	**۰/٤٦**	**۰/۵۲**
S1M2	۷۰/۱۹	۴۲/۲۲	۰/۸۳	۰/۸۹	S1M2	۱۲۴/۵۰	۸۰/۴۸	۰/۴۸	۰/۴۸
S1M3	۶۵/۳۷	۳۹/۱۴	۰/۸۵	۰/۸۹	S1M3	۱۳۷/۳۶	۹۱/۹۱	۰/۴۲	۰/۴۱
S1M4	۶۵/۶۰	۴۰/۳۲	۰/۸۵	۰/۸۹	S1M4	۱۳۲/۵۳	۹۲/۱۱	۰/۴۴	۰/۴۱
S1M5	۶۳/۹۳	۳۹/۴۹	۰/۸۵	۰/۸۹	S1M5	۱۳۷/۵۱	۹۹/۱۷	۰/۴۳	۰/۳۶
S1M6	۶۳/۸۵	۳۹/۱۱	۰/۸۴	۰/۸۹	S1M6	۱۴۱/۱۵	۱۰۴/۸۵	۰/۴۲	۰/۳۳
S2M1	۶۹/۴۸	۴۱/۱۶	۰/۸۵	۰/۸۹	S2M1	۱۱۹/۳۲	۷۹/۵۵	۰/۵۰	۰/۵۲
S2M2	۶۳/۸۷	۳۷/۶۶	۰/۸۵	۰/۹۰	**S2M2**	**۱۱۸/۰۰**	**۷۸/۸۹**	**۰/۵۱**	**۰/۵۳**
S2M3	۶۲/۵۱	۳۶/۵۵	۰/۸۵	۰/۹۰	S2M3	۱۲۰/۳۸	۸۲/۵۲	۰/۵۱	۰/۴۷
S2M4	۶۲/۴۳	۳۸/۰۱	۰/۸۵	۰/۹۰	S2M4	۱۲۲/۳۷	۸۳/۵۸	۰/۵۰	۰/۴۶
S2M5	۶۲/۵۹	۳۷/۳۹	۰/۸۵	۰/۹۰	S2M5	۱۲۶/۶۳	۹۲/۰۷	۰/۴۸	۰/۴۱
S2M6	۶۲/۲۱	۳۷/۲۷	۰/۸۴	۰/۹۰	S2M6	۱۳۲/۰۰	۹۵/۰۰	۰/۴۷	۰/۳۹
S3M1	۶۷/۵۲	۳۹/۱۷	۰/۸۸	۰/۸۹	**S3M1**	**۱۱۳/۰۲**	**٦٧/٢٧**	**۰/٥٤**	**۰/۵۷**
S3M2	۵۵/۶۷	۳۱/۷۲	۰/۸۸	۰/۹۱	S3M2	۱۲۴/۷۲	۷۴/۶۱	۰/۴۹	۰/۵۲
S3M3	۵۶/۸۷	۳۲/۷۴	۰/۸۷	۰/۹۱	S3M3	۱۲۵/۹۱	۷۵/۹۳	۰/۴۸	۰/۵۱
S3M4	۵۷/۲۴	۳۴/۱۱	۰/۸۷	۰/۹۱	S3M4	۱۱۷/۱۵	۷۸/۰۱	۰/۵۲	۰/۵۰
S3M5	۵۹/۰۱	۳۵/۲۴	۰/۸۶	۰/۹۰	S3M5	۱۱۸/۸۷	۸۲/۷۷	۰/۵۲	۰/۴۷
S3M6	۵۸/۵۶	۳۴/۸۶	۰/۸۶	۰/۹۱	S3M6	۱۲۵/۸۹	۸۹/۰۴	۰/۴۹	۰/۴۳
S4M1	۶۰/۴۸	۳۴/۹۰	۰/۸۸	۰/۹۱	S4M1	۱۱۷/۷۱	۸۰/۷۲	۰/۵۲	۰/۵۲
S4M2	۵۵/۵۲	۳۲/۰۲	۰/۸۷	۰/۹۱	**S4M2**	**۱۱۱/۰۰**	**٧٦/٣٦**	**۰/۵۵**	**۰/۵۱**
S4M3	۵۴/۷۶	۳۱/۴۰	۰/۸۷	۰/۹۱	S4M3	۱۱۶/۴۵	۸۱/۰۳	۰/۵۲	۰/۴۸
S4M4	۵۷/۱۲	۳۳/۶۰	۰/۸۷	۰/۹۱	S4M4	۱۱۲/۰۸	۷۹/۹۰	۰/۵۵	۰/۴۹
S4M5	۵۷/۱۹	۳۳/۷۶	۰/۸۷	۰/۹۱	S4M5	۱۲۰/۰۱	۸۷/۲۵	۰/۵۲	۰/۴۴
S4M6	۵۷/۶۹	۳۴/۲۹	۰/۸۶	۰/۹۱	S4M6	۱۲۷/۷۱	۹۲/۷۰	۰/۴۸	۰/۴۱

** مقادیر پررنگ نشان دهنده الگوی برتر در هر سناریو می‌باشد.

جدول ٤- نتایج شاخص‌های ارزیابی عملکرد مدل GPR در مرحله آموزش و آزمون

مرحله آموزش					مرحله آزمون				
الگو	RMSE (m^3/s)	MAE (m^3/s)	KGE	WI	الگو	RMSE (m^3/s)	MAE (m^3/s)	KGE	WI
S1M1	۱۰۸/۷۷	۶۹/۴۸	۰/۷۶	۰/۸۱	**S1M1**	**۹۷/٤۹**	**۷۱/۳۷**	**۰/٥٦**	**۰/٥٤**
S1M2	۷۰/۱۹	۴۲/۲۲	۰/۸۳	۰/۸۹	S1M2	۱۰۴/۳۴	۷۴/۶۸	۰/۵۷	۰/۵۲
S1M3	۶۵/۳۷	۳۹/۱۴	۰/۸۵	۰/۸۹	S1M3	۱۱۵/۱۱	۸۳/۳۹	۰/۵۲	۰/۴۷
S1M4	۶۵/۶۰	۴۰/۳۲	۰/۸۵	۰/۸۹	S1M4	۱۱۷/۳۹	۸۸/۰۰	۰/۵۲	۰/۴۴
S1M5	۶۳/۹۳	۳۹/۴۹	۰/۸۵	۰/۸۹	S1M5	۱۲۲/۷۷	۹۶/۵۷	۰/۵۱	۰/۳۸
S1M6	۶۳/۸۵	۳۹/۱۱	۰/۸۴	۰/۸۹	S1M6	۱۲۷/۶۱	۱۰۱/۳۲	۰/۴۹	۰/۳۵
S2M1	۶۹/۴۸	۴۱/۱۶	۰/۸۵	۰/۸۹	**S2M1**	**۱۰۵/۳۹**	**۷۳/٥٦**	**۰/٥٨**	**۰/٥۳**
S2M2	۶۳/۸۷	۳۷/۶۶	۰/۸۵	۰/۹۰	S2M2	۱۱۱/۷۵	۸۰/۹۹	۰/۵۵	۰/۴۸
S2M3	۶۲/۵۱	۳۶/۵۵	۰/۸۵	۰/۹۰	S2M3	۱۱۰/۱۶	۸۰/۱۵	۰/۵۶	۰/۴۹
S2M4	۶۲/۴۳	۳۸/۰۱	۰/۸۵	۰/۹۰	S2M4	۱۱۸/۰۸	۸۶/۲۸	۰/۵۲	۰/۴۵
S2M5	۶۲/۵۹	۳۷/۳۹	۰/۸۵	۰/۹۰	S2M5	۱۱۶/۸۰	۹۰/۳۶	۰/۵۳	۰/۴۲
S2M6	۶۲/۲۱	۳۷/۲۷	۰/۸۵	۰/۹۰	S2M6	۱۱۱/۹۹	۸۸/۲۵	۰/۵۵	۰/۴۳
S3M1	۶۷/۵۲	۳۹/۱۷	۰/۸۸	۰/۸۹	S3M1	۱۰۹/۶۷	۷۸/۳۲	۰/۵۵	۰/۵۰
S3M2	۵۵/۶۷	۳۱/۷۲	۰/۸۸	۰/۹۱	S3M2	۱۰۸/۳۱	۷۸/۲۵	۰/۵۶	۰/۵۰
S3M3	۵۶/۸۷	۳۲/۷۴	۰/۸۷	۰/۹۱	S3M3	۱۰۷/۸۴	۷۵/۲۴	۰/۵۶	۰/۵۲
S3M4	۵۷/۲۴	۳۴/۱۱	۰/۸۷	۰/۹۱	**S3M4**	**۱۰۷/۵۲**	**۷٤/۹۱**	**۰/٥٦**	**۰/٥۲**
S3M5	۵۹/۰۱	۳۵/۲۴	۰/۸۶	۰/۹۰	S3M5	۱۰۷/۸۷	۷۵/۱۰	۰/۵۶	۰/۵۲
S3M6	۵۸/۵۶	۳۴/۸۶	۰/۸۶	۰/۹۱	S3M6	۱۰۸/۹۳	۷۸/۳۰	۰/۵۵	۰/۵۲
S4M1	۶۰/۴۸	۳۴/۹۰	۰/۸۸	۰/۹۱	**S4M1**	**۱۰۸/۸۷**	**۷۷/۲٤**	**۰/٥٥**	**۰/٥۰**
S4M2	۵۵/۵۲	۳۲/۰۲	۰/۸۷	۰/۹۱	S4M2	۱۰۹/۵۷	۷۷/۹۲	۰/۵۵	۰/۵۱
S4M3	۵۴/۷۶	۳۱/۴۰	۰/۸۷	۰/۹۱	S4M3	۱۲۴/۷۷	۹۰/۴۹	۰/۴۹	۰/۴۲
S4M4	۵۷/۱۲	۳۳/۶۰	۰/۸۷	۰/۹۱	S4M4	۱۰۹/۶۰	۷۷/۶۱	۰/۵۵	۰/۵۰
S4M5	۵۷/۱۹	۳۳/۷۵	۰/۸۷	۰/۹۱	S4M5	۱۱۴/۰۶	۷۸/۸۳	۰/۵۳	۰/۴۹
S4M6	۵۷/۶۹	۳۴/۲۹	۰/۸۶	۰/۹۱	S4M6	۱۱۶/۲۱	۷۹/۲۴	۰/۵۲	۰/۴۹

** مقادیر پررنگ نشان دهنده الگوی برتر در هر سناریو می‌باشد.

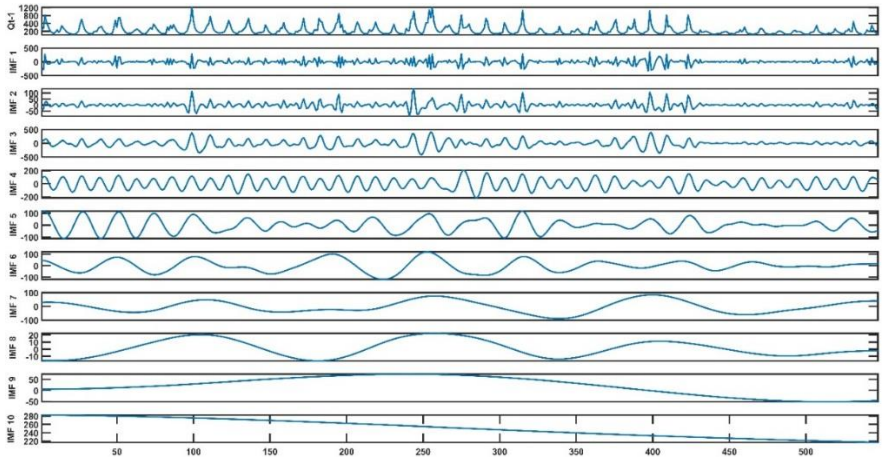
توسعه مدل‌های تلفیقی مبتنی بر تجزیه داده‌ها با استفاده از روش CEEMD

همانگونه که اشاره شد، مدل‌های منفرد از توانایی لازم برای استفاده کامل و بهینه از اطلاعات در دسترس برخوردار نیستند. بدین منظور در این مطالعه بهترین الگوهای ورودی از هر سناریو که با آنها مدل‌های RF و GPR به بهینه‌ترین عملکرد خود می‌رسیدند انتخاب شدند. در گام بعد ورودی‌ها با استفاده از روش CEEMD تجزیه شده و سپس به مدل‌ها معرفی گردیدند. روش CEEMD بر پایه تجزیه سری داده‌ها به IMFهای مختلف و یک سری باقیمانده می‌باشد. در نهایت مجموع IMFها و سری باقیمانده باید بتواند سری تجزیه شده اصلی را بازسازی نماید (روشنگر و قاسم پور، ۱۳۹۹). ایجاد IMFها براساس تفریق تابع پایه از سری داده اصلی است. این مرحله تا زمانی ادامه می‌یابد که تقریباً سری باقیمانده ثابت شود. شکل ۴، نمونه‌ای از نتایج روش CEEMD را در تجزیه داده‌ها نشان می‌دهد. در این شکل، دبی جریان با یک تاخیر (الگوی S1M1) به ۱۰ تابع مد ذاتی یا IMF تجزیه شده است. IMF10 همان سری باقیمانده بوده و نشان دهنده روند کلی در داده‌ها می‌باشد (روشنگر و قاسم پور، ۱۳۹۹). IMFهای یک تا نه نیز زیرسری‌هایی با فرکانس زیاد به کم هستند که هر کدام نشانگر اجزایی متناوب با دوره‌های مشخص می‌باشد (روشنگر و قاسم پور، ۱۳۹۹).

همانگونه که مشاهده گردید خطای مدل‌های منفرد بسیار زیاد بوده و نیاز است تا با ایجاد تغییراتی در ورودی‌ها بتوان اطلاعات موثر را در اختیار این مدل‌ها قرار داد. بدین منظور داده‌های ورودی با استفاده از روش تجزیه مد تجربی یکپارچه کامل تجزیه و به مدل‌ها معرفی گردید. نتایج تحلیل‌های آماری حاصل از کاربرد این روش در جدول ۵ ارائه شده است. با توجه به این جدول مشاهده می‌شود که در سناریوی اول مقدار شاخص RMSE از ۱۱۷/۸۵ متر مکعب بر ثانیه به ۸۷/۱۳ متر مکعب بر ثانیه کاهش یافته است. کاهش خطا با بهبود عملکرد مدل RF نیز همخوانی داشته و شاخص KGE بیش از ۰/۱۳ بهبود یافته است. این امر نشان دهنده تاثیر مثبت پیش‌پردازش داده‌ها

در فرآیند مدل‌سازی می‌باشد.

از دیگر اطلاعاتی که می‌توان از جدول ۵ استنباط نمود اشاره به تفاوتی است که در عملکرد مدل ترکیبی CEEMD-RF با الگوهای مختلف ورودی وجود دارد. به عنوان مثال، روش CEEMD-RF با الگوی تعریف شده سناریوی چهارم با در نظر گرفتن اثر خاصیت تناوبی به بهینه‌ترین دقت دست یافته است. این امر نشان می دهد که خاصیت تناوبی یکی از ورودی‌های مهمی است که می‌تواند بدون آنکه هزینه‌ای را از نظر تامین اطلاعات به فرآیند مدل‌سازی تحمیل نماید، موجبات بهبود عملکرد را فراهم آورد.



شکل ٤- IMFهای حاصل از روش تجزیه مد تجربی یکپارچه کامل برای سری دبی با یک تاخیر

جدول ۵- نتایج شاخص‌های ارزیابی برای عملکرد مدل GPR در مرحله آموزش و آزمون

مدل ترکیبی	مرحله آموزش					مرحله آزمون				
	الگو	RMSE (m^3/s)	MAE (m^3/s)	KGE	WI	الگو	RMSE (m^3/s)	MAE (m^3/s)	KGE	WI
CEEMD-RF	S1M1	۶۹/۸۲	۴۵/۵۳	۰/۸۰	۰/۸۸	S1M1	۸۷/۱۳	۶۹/۱۱	۰/۵۹	۰/۵۶
	S2M2	۴۴/۸۲	۲۸/۴۸	۰/۸۶	۰/۹۲	S2M2	۷۵/۷۷	۶۲/۵۳	۰/۷۰	۰/۶۰
	S3M1	۴۹/۱۰	۲۹/۱۳	۰/۸۸	۰/۹۲	S3M1	۷۸/۳۸	۶۱/۴۸	۰/۶۹	۰/۶۱
	S4M2	۴۲/۴۱	۲۶/۳۱	۰/۸۸	۰/۹۳	S4M2	۷۴/۶۰	۶۱/۹۸	۰/۷۱	۰/۶۰
CEEMD-GPR	S1M1	۱۲۵/۹۰	۸۶/۰۴	۰/۷۵	۰/۷۷	S1M1	۸۶/۵۷	۶۵/۴۷	۰/۵۹	۰/۵۸
	S2M1	۸۷/۶۰	۵۶/۸۰	۰/۸۵	۰/۸۵	S2M1	۷۲/۹۶	۵۵/۷۲	۰/۷۴	۰/۶۴
	S3M4	۵۴/۰۶	۳۸/۷۲	۰/۹۵	۰/۸۹	S3M4	۶۵/۷۸	۴۲/۹۶	۰/۷۵	۰/۷۲
	S4M1	۷۱/۵۳	۴۸/۹۰	۰/۹۰	۰/۸۷	S4M1	۶۹/۶۴	۵۰/۸۴	۰/۷۵	۰/۶۷

بهبود دقت برآورد جریان ورودی به سد دز برای مدل CEEMD-GPR نیز مشهود است (جدول ۵). با توجه به نتایج به دست آمده می‌توان نتیجه گرفت که مدل ترکیبی CEEMD-GPR توانسته است ضمن کاهش خطای تخمین، کارایی برآوردها را نیز بر اساس شاخص KGE به طور متوسط از ۰/۵۲ به ۰/۷۱ ارتقا دهد. بررسی و مقایسه مقادیر شاخص RMSE در مرحله آزمون برای مدل منفرد و ترکیبی GPR نشان می‌دهد که تک تک مقادیر پیش‌بینی شده در مدل تلفیقی تا حدود زیادی با مقادیر اندازه گیری شده مطابقت دارد. مقدار این شاخص آماری برای بهترین الگوهای مختلف ورودی در مدل GPR بین ۹۷/۴۹ تا ۱۰۸/۸۷ متر مکعب در ثانیه تغییر می‌کند، در حالی که این مقدار برای مدل CEEMD-GPR بین ۸۶/۵۷ تا ۶۵/۷۸ متغیر است و بنابراین می‌توان نتیجه گرفت که مدل ترکیبی توسعه داده شده در مقایسه با مدل منفرد GPR توانسته است با ورودی‌های مختلف مقادیر مشاهداتی را به صورت نظیر به نظیر بهتر برآورد نماید.

جدول ۵ نکته مهم دیگری را نیز نشان می‌دهد و آن عملکرد برتر مدل CEEMD-GPR در مقابل CEEMD-RF می‌باشد. مدل تلفیقی GPR در مقایسه با روش تلفیقی RF توانسته است با خطایی کمتر بهینه‌ترین مقادیر را برای جریان ورودی به سد دز برآورد نماید. روش CEEMD-RF با الگوی ورودی S4M2 (شامل پارامترهای جریان با دو تاخیر، بارش با دو تاخیر ماهانه و ترم تناوبی) به حداکثر دقت رسیده و براساس شاخص RMSE، ۳۶/۴۰ متر مکعب بر ثانیه

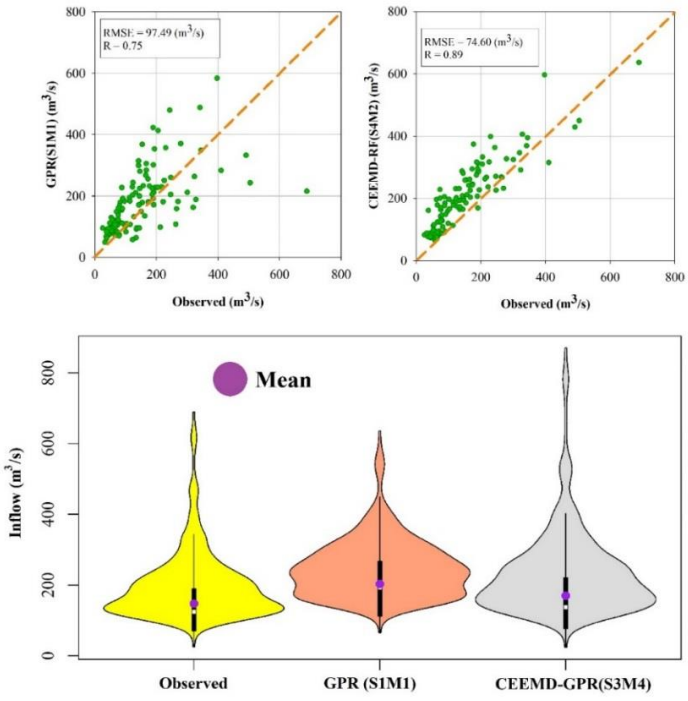
خطا را کاهش داده است. این در حالی است که مدل CEEMD-GPR با الگوی S3M4 (شامل پارامترهای جریان با چهار تاخیر، بارش با چهار تاخیر ماهانه و ترم تناوبی) به بهترین عملکرد خود دست یافته و شاخص RMSE را در الگوی فوق از ۱۰۷/۵۲ متر مکعب در ثانیه به ۶۵/۷۸ متر مکعب در ثانیه (بیش از ۴۰ متر مکعب بر ثانیه بهبود خطا) کاهش داده است.

استفاده از شاخص‌های عددی نمی‌تواند اطلاعات لازم در خصوص توزیع داده‌ها را در اختیار کاربران قرار دهد. از این رو استفاده از نمودارهایی که بتواند توزیع داده‌ها را نشان داده و تفاوت مدل‌ها را نشان دهد بسیار حائز اهمیت است. از جمله این نمودارها می‌توان به نمودار جعبه‌ای، نمودار پراکندگی و نمودار ویولونی اشاره نمود. نمودار جعبه‌ای صرفا اطلاعاتی را در خصوص میانگین، چارک‌ها و بیشینه و کمینه داده‌ها در اختیار قرار داده و نمودار پراکندگی نیز صرفاً دوری و نزدیکی به خط یک به یک را نشان می‌دهد. در این بین نمودار ویلونی علاوه بر میانگین، بیشینه و کمینه داده‌ها، توزیع آنها را نیز مورد بررسی قرار داده و تطابق مقادیر حاصل از مدل‌ها را از نظر بزرگی و کوچکی با مقادیر مشاهداتی نشان می‌دهد. در شکل‌های ۵ و ۶ نمودارهای پراکندگی و ویلونی حاصل از مدل‌های RF، GPR، CEEMD-RF و CEEMD-GPR با بهترین الگوهای ورودی ارائه شده است.

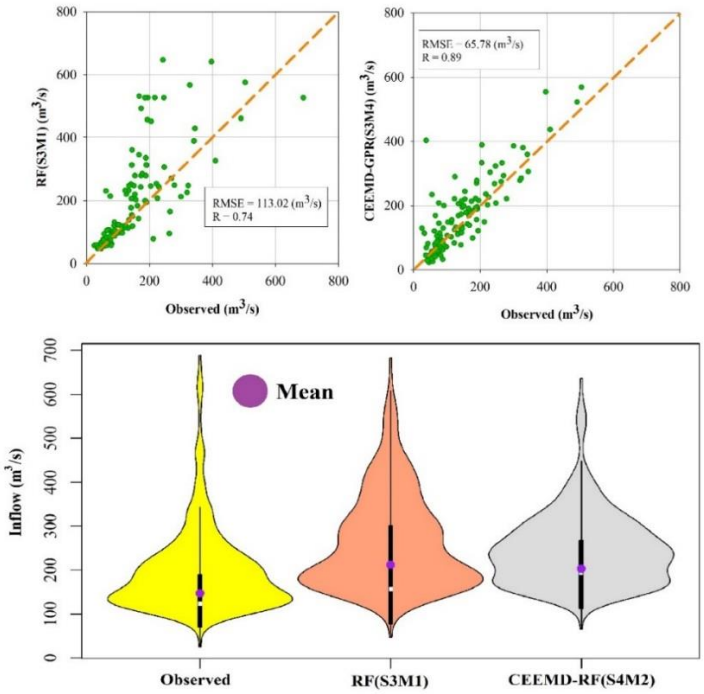
با توجه به شکل‌های فوق مشاهده می‌شود که مدل هیبریدی CEEMD-GPR در مقایسه با حالت منفرد، دچار بیش‌برازش شده اما توانسته است مقادیر کم و متوسط را به خوبی برآورد نموده و در نتیجه کارایی و خطای مدل‌سازی را به طور قابل ملاحضه‌ای کاهش دهد. مدل GPR در مقادیر بیشینه دچار کم‌برآورد شده و مقادیر کمینه و نزدیک به میانگین را بیش از مقدار مشاهداتی برآورد نموده و در نتیجه میانگین مقادیر تخمینی GPR در مقایسه با مقادیر مشاهداتی، بیشتر می‌باشد (به نقطه بنفش وسط ویولون‌ها توجه شود). برای مدل CEEMD-RF نیز مقادیر بیشینه کمتر از واقعیت تخمین زده شده اما ویولون آن به مقادیر مشاهداتی شباهت بیشتری دارد. مدل RF برای مقادیر بزرگ، بیش‌برآورد داشته و در نتیجه انتهای ویولون‌ها در قسمت اعداد بزرگ نسبت به ویولون مشاهداتی ضخیم‌تر می‌باشد. این امر باعث شده میانگین مقادیر مشاهداتی با مقادیر تخمینی اختلاف فاحشی را نشان دهد.

نتیجه گیری

در این پژوهش سعی گردید مدل‌های جنگل‌های تصادفی (RF) و رگرسیون فرآیند گاوسی (GPR) با روش پیش‌پردازش تجزیه مد تجربی یکپارچه کامل (CEEMD) تلفیق شده و عملکرد آنها در دوحالت منفرد و هیبریدی برای برآورد جریان ورودی به سد دز مورد ارزیابی قرار گیرد. نتایج حاصل از پژوهش حاضر به شرح زیر قابل ارائه می‌باشد:



شکل ۵- نمودارهای پراکندگی و ویولونی حاصل از مدل های GPR و CEEMD-GPR با بهترین الگوهای ورودی



شکل ٦- نمودارهای پراکندگی و ویولونی حاصل از مدل های RF و CEEMD-RF با بهترین الگوهای ورودی

- نتایج به دست آمده نشان داد که مدل GPR با استفاده از تابع کرنل RBF نسبت به توابع چند جمله‌ای و خطی، در پیش‌بینی جریان ماهانه وروی به سد دز بهترین عملکرد را به خود اختصاص داده است.
- در حالتی که الگوهای ورودی به مدل‌های منفرد با ترم تناوبی همراه باشد، نتایج نشان داد که استفاده از این ضرایب می‌تواند نسبت به حالت ساده دقت برآوردها را بهبود بخشد. ترم پریودیک صرفاً براساس یک رابطه ریاضی ساده یا با توجه به شماره ماه‌ها محاسبه شده و هیچ‌گونه زمان و هزینه مازادی را از لحاظ جمع‌آوری داده‌ها به پژوهش‌گران تحمیل نمی‌کند. اما در بسیاری از مطالعات سعی می‌گردد برای بهبود دقت و عملکرد مدل‌های هوشمند از داده‌های هیدرولوژیکی دیگر نظیر بارش، دماهای مینیمم، متوسط و ماکزیمم، تبخیر و غیره بهره برده می‌شود. در این مطالعه از داده‌های بارش نیز در ورودی-های مدل بهره گرفته شد که همراهی آن با ترم تناوبی عملکرد مدل را به طور قابل ملاحضه‌ای بهبود بخشید. در این پژوهش ترم تناوبی یک حالت ساده خطی داشته و بنابراین پیشنهاد می-شود در مطالعات دیگر از حالت غیرخطی آن در مدل‌سازی فرآیندهای هیدرولوژیک بهره گرفته شود.
- در این مطالعه، علاوه بر روش‌های ساده ذکر شده از یک روش ریاضی پیچیده مبتنی بر تجزیه سیگنال تحت عنوان روش تجزیه مد یکپارچه کامل برای بهبود عملکرد مدل‌های RF و GPR استفاده شد. نتایج نشان داد که استفاده از این روش خطای مدل‌های منفرد RF و GPR را به ترتیب حدود ۵۲ و ۴۹ درصد کاهش داده و به خوبی می‌تواند میانگین و واریانس داده-ها را مدل‌سازی نماید. در این بین بهترین عملکرد متعلق به مدل CEEMD-GPR با الگوی ورودی S3M4 بوده و مقادیر آمارهای RMSE، MAE و KGE در مرحله آزمون به ترتیب برابر با ۶۵/۷۸ متر مکعب در ثانیه، (m³/s) ۴۲/۹۶ و ۰/۷۵ محاسبه گردید.

منابع

ثاقبیان، س. م. ۱۳۹۹. پیش بینی زمانی و مکانی دبی جریان با استفاده از روش های تلفیقی هوش مصنوعی و پیش پردازش و پس پردازش سری زمانی. نشریه آبیاری و زهکشی ایران، ۱۴(۴): ۱۱۳۷-۱۱۵۱.

منتصری، م. و زمان زاد قویدل، س. ۱۳۹۵. مقایسه عملکرد مدل های هوش مصنوعی در تخمین پارامترهای کیفی آب رودخانه در دوره های کم آبی و پرآبی. آب و خاک، ۳۰(۶): ۱۷۳۳-۱۷۴۷.

رضازاده جودی، ع. و ستاری، م. ت. ۱۳۹۵. ارزیابی عملکرد روش های مبتنی بر کرنل در تخمین میزان بار رسوبی معلق رودخانه (مطالعهٔ موردی: رودخانهٔ صوفی‌چای مراغه). پژوهش های جغرافیای طبیعی، ۴۸(۳): ۴۱۳-۴۲۹.

احمدی، ف. ۱۳۹۹. ارزیابی عملکرد روش‌های ماشین‌بردار پشتیبان و سیستم استنتاج عصبی فازی تطبیقی در پیش‌بینی جریان ماهانه رودخانه‌ها (مطالعه موردی رودخانه‌های نازلو و سزار). تحقیقات آب و خاک ایران، ۵۱(۳): ۶۷۳-۶۸۳.

روشنگر، ک. و قاسم پور، ر. ۱۳۹۹. بهبود پیش بینی بارش ماهانه با استفاده از مدل تلفیقی بر پایه روش کرنل- تبدیل موجک و تجزیه ی یکپارچه مد تجربی کامل .نشریه مهندسی عمران امیرکبیر، ۵۲(۱۰): ۲۶۴۹-۲۶۶۰.

Abda, Z., Zerouali, B., Chettih, M., Guimaraes Santos, C.A., de Farias, C.A.S. and Elbeltagi, A. 2022. Assessing machine learning models for streamflow estimation: a case study in Oued Sebaou watershed (Northern Algeria). Hydrological Sciences Journal. 67(9): 1328-1341.

Ahmadi, F., Mehdizadeh, S. and Nourani, V. 2022. Improving the performance of random forest for estimating monthly reservoir inflow via complete ensemble empirical mode decomposition and wavelet analysis. Stochastic Environmental Research and Risk Assessment. 1-16.

Al-Abadi, A. M., and Shahid, S. 2016. Spatial mapping of artesian zone at Iraqi southern desert using a GIS-based random forest machine learning model. Modeling Earth Systems and Environment. 2: 1-17.

Ali, M., Prasad, R., Xiang, Y. and Yaseen, Z.M. 2020. Complete ensemble empirical mode decomposition hybridized with random forest and kernel ridge regression model for monthly rainfall forecasts. Journal of Hydrology. 584: 624-647.

Beven, K. 2020. Deep learning, hydrological processes and the uniqueness of place. Hydrological Processes. 34(16): 3608-3613.

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

Darbandsari, P. and Coulibaly, P. 2020. Introducing entropy-based Bayesian model averaging for streamflow forecast. Journal of Hydrology. 591: 125-177.

Elbeltagi, A., Pande, C.B., Kumar, M., Tolche, A.D., Singh, S.K., Kumar, A. and Vishwakarma, D.K., 2023. Prediction of meteorological drought and standardized precipitation index based on the random forest (RF), random tree (RT), and Gaussian process regression (GPR) models. Environmental Science and Pollution Research. 6: 1-20.

Fang, W., Huang, S., Ren, K., Huang, Q., Huang, G., Cheng, G. and Li, K. 2019. Examining the applicability of different sampling techniques in the development of decomposition-based streamflow forecasting models. Journal of Hydrology. 568: 534-550.

Fathabadi, A., Seyedian, S.M., and Malekian, A. 2022. Comparison of Bayesian, k-Nearest Neighbor and Gaussian process regression methods for quantifying uncertainty of suspended sediment concentration prediction. Science of the Total Environment. 818: 151-160.

Hosseinzadeh, P., Nassar, A. Boubrahimi S.F. and Hamdi S.M. 2023. ML-Based Streamflow Prediction in the Upper Colorado River Basin Using Climate Variables Time Series Data. Hydrology. 10(2): 29-45.

Huang, N.E., Shen, Z., Long, S.R., Wu, M.C., Shih, H.H., Zheng, Q., Yen, N.C., Tung, C.C. and Liu, H.H. 1998. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences. 454(1971): 903-995.

Huang, N., Lu, G. and Xu, D. 2016. A permutation importance-based feature selection method for short-term electricity load forecasting using random forest. Energies. 9(10): 767.

Ikram, R.M.A., Goliatt, L., Kisi, O., Trajkovic, S. and Shahid, S. 2022. Covariance matrix adaptation evolution strategy for improving machine learning approaches in streamflow prediction. Mathematics. 10(16): 49-71.

Katipoğlu, O. M. 2023. Prediction of streamflow drought index for short-term hydrological drought in the semi-arid Yesilirmak Basin using Wavelet transform and artificial intelligence techniques. Sustainability. 15(2): 110-119.

Keshvari R. Imani M. and Parsa Moghaddam M. 2022. Short Term Load Forecasting Using Empirical Mode Decomposition, Wavelet Transform and Support Vector Regression. Signal and Data Processing. 19(3): 35-48.

Latifoğlu, L. 2022. A novel approach for prediction of daily streamflow discharge data using correlation based feature selection and random forest method. International Advanced Researches and Engineering Journal 6 (1): 1-7.

Li, X., Sha, J. and Wang, Z.L. 2019. Comparison of daily streamflow forecasts using extreme learning machines and the random forest method. Hydrological Sciences Journal. 64(15): 1857-1866.

Lin, Y., Wang, D., Wang, G., Qiu, J., Long, K., Du, Y., Xie, H., Wei, Z., Shangguan, W. and Dai, Y. 2021. A hybrid deep learning algorithm and its application to streamflow prediction. Journal of Hydrology. 601: 126-136.

Meng, E., Huang, S., Huang, Q., Fang, W., Wang, H., Leng, G., Wang, L. and Liang, H. 2021. A Hybrid

Zhang, H. 2022. Multi-Variables-Driven Model Based on Random Forest and Gaussian Process Regression for Monthly Streamflow Forecasting. Water. 14(11): 18-28.

Torres, M.E., Colominas, M.A., Schlotthauer, G. and Flandrin, P. 2011. A complete ensemble empirical mode decomposition with adaptive noise. In 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4144-4147). IEEE.

Wang, J. 2020. An intuitive tutorial to Gaussian processes regression. ArXiv preprint arXiv: 2009.10862.

Were, K., Bui, D.T., Dick, B. and Singh, B.R. 2015. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. Ecological Indicators. 52: 394-403.

Willmott, C.J., Robeson, S.M. and Matsuura, K. 2012. A refined index of model performance. International Journal of Climatology. 32(13): 2088-2094.

Wu, Z. and Huang, N.E. 2009. Ensemble empirical mode decomposition: a noise-assisted data analysis method. Advances in Adaptive Data Analysis. 1(01): 1-41.

Zhu, S., Luo, X., Xu, Z. and Ye, L. 2019. Seasonal streamflow forecasts using mixture-kernel GPR and advanced methods of input variable selection. Hydrology Research. 50(1): 200-214.

VMD-SVM model for practical streamflow prediction using an innovative input selection framework. Water Resources Management. 35(4): 1321-1337.

Niu, W. J. and Feng, Z. K. 2021. Evaluating the performances of several artificial intelligence methods in forecasting daily streamflow time series for sustainable water resources management. Sustainable Cities and Society. 64: 102-122.

Orellana-Alvear, J., Célleri, R., Rollenbeck, R., Muñoz, P., Contreras, P. and Bendix, J. 2020. Assessment of native radar reflectivity and radar rainfall estimates for discharge forecasting in mountain catchments with a random forest model. Remote Sensing 12(12): 19-86.

Saray, M.H., Eslamian, S.S., Klöve, B. and Gohari, A. 2020. Regionalization of potential evapotranspiration using a modified region of influence. Theoretical and Applied Climatology. 140(1): 115-127.

Shannon, C.E. 2001. A mathematical theory of communication. ACM SIGMOBILE Mobile Computing and Communications Review. 5(1): 3-55.

Shen, Y., Ruijsch, J., Lu, M., Sutanudjaja, E.H. and Karssenberg, D. 2022. Random forests-based error-correction of streamflow from a large-scale hydrological model: Using model state variables to estimate error terms. Computers & Geosciences. 159: 105019.

Sun, N., Zhang, S., Peng, T., Zhang, N., Zhou, J. and

Iranian Journal of Irrigation and Drainage
No.4, Vol. 17, Oct.-Nov. 2023, p. 685-701

نشریه آبیاری و زهکشی ایران
شماره۴، جلد ۱۷، مهر - آبان ۱۴۰۲، ص. ۷۰۱-۶۸۵

Development of Integrated Machine Learning Models Based on Complete Ensemble Empirical Mode Decomposition Method for Estimating Dam Inflow (Case study: Dez Dam)

N. Mousazadeh[1], A. M. Akhoond-Ali[2], F. Ahmadi[3]*
Recived: Mar.07, 2023 Accepted: May.18, 2023

Abstract

Estimating the inflow to the reservoir of dams is of particular importance in planning and optimal management of water resources, water supply needed by different sectors and flood management. Therefore, in the current research, it was tried to evaluate the performance of random forest (RF) and Gaussian process regression (GPR), machine learning models by using the preprocessing method of complete ensemble experimental mode decomposition (CEEMD) for estimating the monthly inflow to Dez Dam in the period of 1971 to 2017. For this purpose, the input patterns in four different scenarios, including the use of flow data with time delays, the combination of flow and precipitation data with time delays, and adding periodic term to the previous two modes, were prepared and introduced to standalone models. The results showed that each model achieves its maximum accuracy with different scenarios, and in the meantime, the performance of the GPR model was the best with an RMSE value of 97.49 (m^3/s). After determining the best input patterns in each scenario, the relevant data were analyzed by CEEMD method and the modeling process was performed with RF and GPR methods. Based on the evaluation criteria, the error reduction and accuracy increase in the developed integrated models were significantly evident. So that the CEEMD-GPR model was able to reduce the value of the RMSE index by about 47 (m3/s). The same behavior was observed for the CEEMD-RF model. In general, the performance of CEEMD-GPR is more suitable compared to all the developed models (single or integrated) and it is recommended for predicting the inflow to Dez Dam.

Keywords: Delay time, Hybrid model, Intrinsic mode function, Standalone model

1 - Msc student in Water Resources Engineering, Faculty of Water and Environmental Engineering, Shahid Chamran University of Ahvaz, Ahvaz, Iran
2 - Professor, Department of Hydrology and Water Resources, Faculty of Water and Environmental Engineering, Shahid Chamran University of Ahvaz, Ahvaz, Iran
3 - Assistant professor, Department of Hydrology and Water Resources, Faculty of Water and Environmental Engineering, Shahid Chamran University of Ahvaz, Ahvaz, Iran
(* - Corresponding Author Email: f.ahmadi@scu.ac.ir).