مقایسه روش‌های برنامه‌ریزی بیان ژن و رگرسیون‌های پارامتریک و ناپارامتریک در پیش‌بینی دبی میانگین روزانه رودخانه کارون (مطالعه موردی: ایستگاه هیدرومتری ملاثانی)

مهدی علی‌نژادی، سیدفرهاد موسوی* و خسرو حسینی[۱]

(تاریخ دریافت: ۱۳۹۸/۱۲/۱۷ ؛ تاریخ پذیرش: ۱۳۹۹/۵/۱۵)

چکیده

امروزه، پیش‌بینی جریان رودخانه‌ها از مباحث مهم در هیدرولوژی و منابع آب است که می‌توان از نتایج الگوبندی جریان رودخانه در مدیریت منابع آب، مدیریت سازه‌های آبی و پیش‌بینی سیل استفاده کرد. در این تحقیق، عملکرد مدل برنامه‌ریزی بیان ژن (GEP)، رگرسیون پارامتریک خطی (LR)، رگرسیون پارامتریک غیرخطی (NLR) و همچنین روش ناپارامتریک K- نزدیک‌ترین همسایگی (K-NN)، در پیش‌بینی میانگین دبی روزانه رودخانه کارون در محل ایستگاه هیدرومتری ملاثانی طی دوره آماری ۹۶-۱۳۴۶ مورد ارزیابی قرار گرفته است. ترکیب‌های مختلفی از داده‌های ثبت شده به‌عنوان الگوی ورودی برای پیش‌بینی دبی جریان استفاده شد. نتایج به‌دست آمده حاکی از عملکرد بهتر مدل برنامه‌ریزی بیان ژن با ضریب تبیین (R^2 =۰/۸۲۷)، جذر میانگین مربعات خطا (RMSE = ۵۹/۴۵) و میانگین خطای مطلق (MAE=۲۶/۶۴) در مرحله صحت‌سنجی برای پیش‌بینی دبی روزانه رودخانه کارون در ایستگاه ملاثانی با تأخیر ۵ روز، در مقایسه با روش‌های LR، NLR و K-NN بوده است. همچنین، ارزیابی عملکرد مدل‌ها در پیش‌بینی مقادیر حداکثر آبدهی جریان نشان داد که همه مدل‌ها میزان جریان را در بیشتر موارد کمتر از مقدار مشاهداتی تخمین زده‌اند.

واژه‌های کلیدی: آبدهی رودخانه، برنامه‌ریزی بیان ژن، رگرسیون خطی و غیرخطی، K – نزدیک‌ترین همسایگی، کارون.

۱. گروه مهندسی آب و سازه‌های هیدرولیکی، دانشکده مهندسی عمران، دانشگاه سمنان

*: مسئول مکاتبات: پست الکترونیکی: fmousavi@semnan.ac.ir

مقدمه

مدل‌های پیش‌بینی رواناب مجموعه‌ای از روش‌های بسیار پرکاربرد در زمینه هیدرولوژی و مدیریت منابع آب هستند که می‌توانند تغییرات هیدرولوژیک برای برنامه‌ریزی منابع آب را در اختیار ما قرار دهند. به‌کارگیری روش‌های مدل‌سازی داده‌مبنا (Data driven modeling, DDM) بخش قابل توجهی از فعالیت‌ها و پژوهش‌های انجام شده در زمینه مدل‌های پیش‌بینی رواناب را به‌خود اختصاص داده است. عملکرد این مدل‌ها مبتنی بر یافتن رابطه‌ای بین داده‌های ورودی و خروجی یک سیستم است که دانش فرد متخصص در مورد فرایندهای فیزیکی مرتبط با سیستم می‌تواند به بهبود انتخاب متغیرها و نتایج حاصل از به‌کارگیری مدل منتهی شود (۳). گووین‌داراجو (۸) الگوهای متداول هیدرولوژیک را به الگوهای ریاضی- فیزیکی، ژئومورفولوژیک و تجربی تقسیم کرده است. الگوهای سری زمانی، در زمره مدل‌های ریاضی- فیزیکی محسوب می‌شوند و توانایی زیادی در الگوبندی پدیده‌های خطی و غیرخطی دارند (۲۵). ماهیت جریان رودخانه اغلب متغیر است. بنابراین، مدل‌های خطی کارایی لازم را در این خصوص نداشته و لازم است از مدل‌های غیرخطی استفاده شود. یکی از این مدل‌ها، برنامه‌ریزی بیان ژن (Gene expression programming, GEP) است (۱۸).

برنامه‌ریزی بیان ژن یک روش پرکاربرد در منابع آب و هیدرولوژی است که در آن، راه حل مسئله با برنامه‌نویسی ارائه می‌شود. الگوریتم‌های تکاملی، که برنامه‌ریزی ژنتیک عضوی از آنها هستند، توانایی الگوبندی فرایندهای کاملاً غیرخطی را دارند (۲۵). بنا به اهمیت موضوع، پژوهشگران متعددی در جهان با استفاده از این روش اقدام به پیش‌بینی دبی رودخانه‌ها کرده‌اند. زورن و شمس‌الدین (۲۶) پتانسیل استفاده از مدل برنامه‌ریزی بیان ژن را برای پیش‌بینی سیلاب، در مقایسه با روش‌های معمول تخمین سیلاب، در منطقه اوکلند کشور نیوزلند مطالعه کرده‌اند که نتایج نشان‌دهنده عملکرد مطلوب این مدل بوده است.

زمانی و همکاران (۲۵) با استفاده از برنامه‌ریزی بیان ژن، سری‌های زمانی خطی و غیرخطی و شبکه‌های عصبی مصنوعی (Artificial neural networks, ANN)، اقدام به پیش‌بینی آبدهی روزانه ایستگاه ارمند روی رودخانه کارون کرده‌اند. نتایج نشان داده که روش برنامه‌ریزی بیان ژن، عملکرد بهتری در مقایسه با سایر روش‌های به‌کار گرفته شده داشته است. در پژوهش سلطانی و همکاران (۲۱) مدل برنامه‌ریزی ژن در مدل‌سازی فرایند بارش- رواناب به‌کار رفته است. با توجه به دقت و سادگی حاصل از مجموعه عملگرهای ریاضی، این مدل به عنوان مدل بارش- رواناب حوضه آبخیز لیقوان پیشنهاد شده است.

مدل‌های مبتنی بر رگرسیون نیز از روش‌هایی هستند که برای پیش‌بینی جریان رودخانه مورد توجه پژوهشگران قرار گرفته‌اند. سان و همکاران (۲۲) با استفاده از روش‌های ANN، فیلتر کالمن و رگرسیون خطی چندگانه (Multi-linear regression, MLR) به پیش‌بینی آبدهی رودخانه باکل در سنگال پرداختند. نتایج نشان داده که رگرسیون خطی چندگانه در گام زمانی کوتاه‌مدت عملکرد مناسب‌تری نسبت به سایر روش‌های مورد استفاده داشته است.

کیم و کالوآراچچی (۱۲) در یکی از زیرحوضه‌های رود نیل واقع در اتیوپی از روش رگرسیون خطی به‌منظور پیش‌بینی رواناب استفاده کرده‌اند که نتایج نشان‌دهنده دقت بالای این روش در پیش‌بینی رواناب بوده است.

خدمتی و همکاران (۱۱) با استفاده از روش رگرسیون چندمتغیره، دبی اوج با دوره بازگشت متفاوت را در حوضه‌های جنوب شرق ایران برآورد کرده‌اند.

از جمله مدل‌های غیرخطی دیگر مرسوم برای پیش‌بینی جریان رودخانه، روش ناپارامتریک نزدیک‌ترین همسایگی (K- nearest neighbor, K-NN) است که همواره مورد توجه پژوهشگران بوده و مطالعات انجام شده حاکی از کارایی زیاد این روش است (۵).

شارما و لال (۲۰) یک روش ناپارامتریک برگرفته از

الگوریتم K-NN را برای شبیه‌سازی بارش روزانه ارائه داده‌اند. روش پیشنهادی آنها برای یک دوره ۱۲۳ ساله بارش روزانه در سیدنی استرالیا مورد بررسی قرار گرفته است و عملکرد آن موفقیت‌آمیز گزارش شده است.

رن و همکاران (۱۷) از روش‌های ترکیبی الگوریتم K-NN برای پیش‌بینی سیلاب حوضه گدونگ (Gedong) در کشور چین استفاده کرده‌اند. نتایج این تحقیق نشان می‌دهد که اگوریتم K-NN توانایی خوبی در برآورد سیلاب حوضه‌های کوچک و متوسط داشته است.

وو و چاو (۲۴) عملکرد مدل‌های ARMA، ANN و K-NN را در پیش‌بینی جریان ماهانه در چندین رودخانه در کشور چین مورد بررسی قرار داده‌اند. نتایج به‌دست آمده حاکی از عملکرد قابل قبول K-NN در مقایسه با سایر روش‌ها بوده است.

در ایران نیز مطالعاتی در خصوص به‌کارگیری روش K-NN در زمینه پیش‌بینی جریان رودخانه انجام شده است. عزمی و عراقی نژاد (۴) از روش K-NN به‌منظور پیش‌بینی جریان رودخانه در حوضه بالادست سد زاینده‌رود استفاده کرده‌اند و این روش را برای سری‌های تاریخی بلندمدت مناسب دانسته اند.

آهنی و شوریان (۳) با استفاده از چهار تکنیک مدل‌سازی داده‌مبنا شامل MLR، K-NN، ANN و ANFIS به پیش‌بینی رواناب زیرحوضه سراب هنده در حوضه آبریز دریاچه نمک پرداختند. نتایج این تحقیق نشان داد که الگوریتم K-NN دارای توانایی بالایی در پیش‌بینی جریان حوضه بوده است.

با توجه به قرار گرفتن کلان‌شهر اهواز، نیروگاه‌های برق زرگان و رامین و سازه‌های آبی متعدد در پایین‌دست ایستگاه هیدرومتری ملاثانی، هرگونه تصمیم‌گیری درخصوص مدیریت منابع آب، مطالعات رودخانه کارون و یا سیستم‌های هشدار سیل در پایین‌دست این ایستگاه، نیازمند پیش‌بینی مناسب میزان آبدهی روزانه آن است که ضرورت مطالعه حاضر را بیشتر نشان می‌دهد. بنابراین، هدف از این تحقیق، پیش‌بینی آبدهی متوسط روزانه ایستگاه ملاثانی با استفاده از مدل برنامه‌ریزی بیان ژن (GEP)، الگوریتم K- نزدیک‌ترین همسایگی (K-NN)، رگرسیون‌های خطی و غیرخطی و مقایسه نتایج آنها است.

مواد و روش‌ها

منطقه مورد مطالعه و داده‌ها

رودخانه کارون، واقع در حوضه آبریز کارون بزرگ (شکل ۱)، در جنوب غربی ایران، به طول حدود ۸۹۰ کیلومتر، یکی از طولانی‌ترین و پرآب‌ترین رودخانه‌های ایران محسوب می‌شود. مساحت حوضه آبریز این رودخانه ۶۶۹۳۰ کیلومتر مربع است. بخش عمده مساحت حوضه (حدود ۴۵۶۳۰ کیلومتر مربع) در مناطق کوهستانی و حدود ۲۱۳۰۰ کیلومتر مربع در منطق دشتی و کوهپایه‌ای واقع شده است (۲).

ایستگاه مورد مطالعه در این تحقیق، ایستگاه هیدرومتری ملاثانی متعلق به وزارت نیرو بوده که در طول جغرافیایی ′۵۳ °۴۸ شرقی، عرض جغرافیایی ′۳۵ °۳۱ شمالی و ارتفاع ۱۸ متر از سطح دریا، در سال ۱۳۴۴ تأسیس شده و در فاصله ۳۵ کیلومتری شمال شهر اهواز واقع شده است (شکل ۱). حوضه بالادست این ایستگاه به مساحت ۵۹۹۸۲ کیلومتر مربع بین طول جغرافیایی ′۳۱ °۴۸ تا ′۵۰ °۵۱ شرقی و عرض جغرافیایی ′۴۴ °۳۳ تا ′۳۵ °۳۱ شمالی قرار دارد (۱۵). این حوضه شامل دو رودخانه بزرگ دز و کارون است که سد دز روی شاخه دز و سدهای کارون ۳، شهید عباسپور و گتوند علیا از عمده‌ترین سدها روی شاخه کارون آن هستند. داده‌های مورد استفاده در این تحقیق، میانگین آبدهی روزانه ۵۱ ساله ایستگاه هیدرومتری ملاثانی در فاصله سال‌های ۹۶-۱۳۴۶ است که از سازمان آب و برق استان خوزستان به‌دست آمده‌اند.

روند داده‌های سری زمانی

پارامترها و داده‌های هیدرولوژیک بایستی جنبه تصادفی داشته و فاقد روند (Trend) باشند. درصورتی که سری زمانی داده‌های هیدرولوژیک به‌شکل یکنواخت سیر صعودی یا نزولی داشته باشند در این حالت گفته می‌شود که داده‌ها دارای روند

شکل ۱. موقعیت حوضه کارون بزرگ، ایستگاه هیدرومتری ملاثانی و سدهای بالادست آن

بوده و ضرورت دارد که روند آنها برطرف شود. یکی از راه‌کارهای بررسی روند داده‌ها، استفاده از آزمون من- کندال (Mann-Kendall) است (۹). این تست تحت فرض H_0 داده‌ها از سری که مستقل و دارای توزیع یکسان هستند گرفته شده‌اند. اگر x_j و x_k مقادیر سری زمانی در زمان‌های مشاهداتی j^{th} و k^{th} از میان n داده مشاهداتی باشند مقدار S که جمع آماری کندال نام دارد از رابطه ۱ حاصل می‌شود:

$$S = \sum_{K=1}^{n-1} \sum_{j=k+1}^{n} \mathrm{sgn}(x_j - x_k) \qquad (۱)$$

یکی دیگر از پارامترهای آماری مورد نیاز آزمون من- کندال، مقدار استاندارد Z (رابطه ۲) است:

$$Z = \begin{cases} (S-1)/\sqrt{V(S)} & \text{for } S > 0 \\ 0 \text{ for } S = 0 \\ (S+1)/\sqrt{V(S)} & \text{for } S < 0 \end{cases} \qquad (۲)$$

مقدار V(S) رابطه ۲ از رابطه ۳ حاصل می‌شود:

$$V(S) = \frac{n(n-1)(2n+5)}{18} \qquad (۳)$$

میزان Z در سطح ۵ درصد آزمون می‌شود. اگر Z کوچک‌تر از ۱/۹۶+ و بزرگ‌تر از ۱/۹۶- باشد، داده‌ها فاقد روند بوده و در سطح اعتماد ۹۵ درصد تصادفی‌اند (۸).

آزمون سیر تناوبی

فرمول‌بندی یک مدل ریاضی و ارزیابی یک سری زمانی توسط آن، شامل بررسی سیر تناوبی داده‌های مورد استفاده است. سری‌های زمانی حاصل از برداشت‌های طبیعی، بیشتر نوعی رفتار تناوبی از خود نشان می‌دهند. بنابراین، برای تجزیه و تحلیل آنها باید فاکتور تناوبی مشخص شده و درصورت وجود سیر تناوبی، بایستی از سری زمانی حذف شوند. روش‌های گوناگونی برای شناسایی سیر تناوبی داده‌ها وجود دارد نظیر: الف) رسم نمودار زمانی داده‌ها، ب) آزمون ترسیم همبستگی-نگار (Correlogram Test) سری داده‌ها و ج) آزمون ترسیم دوره‌نگار (Periodogram Test) (۱۴).

آزمون کفایت داده‌ها

یکی از روش‌های آزمودن کفایت طول داده‌های سری‌های

زمانی برای استفاده در مدل‌سازی، استفاده از ضریب هرست (Hurst coefficient) است. این ضریب برای سنجش حافظه بلندمدت یک سری زمانی استفاده می‌شود. حافظه سری زمانی بر اساس مشاهده رویدادهای حدی آن در یک بازه معین از سری تعریف می‌شود. ضریب هرست از رابطه ۴ به دست می آید:

$$K = \frac{\text{Log}\frac{R}{\sigma}}{\text{Log}\frac{N}{2}} \qquad (۴)$$

که N تعداد داده در سری زمانی و σ انحراف معیار سری است. در رابطه ۴، R برابر است با تفاوت بین مقادیر مثبت و منفی انحراف از میانگین سری زمانی که به‌صورت تجمعی محاسبه شده باشد:

$$R = S+ - S \qquad (۵)$$

مقدار ضریب هرست برابر با ۰/۵ برای سری زمانی نشان‌دهنده یک سری مستقل نرمال است. هرچه این ضریب از ۰/۵ بیشتر باشد، حافظه بلندمدت در سری زمانی بیشتر است و نیازی برای طولانی کردن اطلاعات سری زمانی نیست (۱۰).

برنامه‌ریزی بیان ژن

روش برنامه‌ریزی بیان ژن (GEP)، ترکیب و توسعه یافته الگوریتم ژنتیک (GA) و برنامه‌ریزی ژنتیک (GP) است که توسط فریرا (۷) ابداع شده است. در این روش، کروموزوم‌های خطی و ساده با طول ثابت، مشابه با الگوریتم ژنتیک و ساختارهای شاخه‌ای با اندازه و شکل‌های متفاوت، مشابه با درختان تجزیه در برنامه‌ریزی ژنتیک، ترکیب می‌شوند. مرحله نخست در GEP، تشکیل جمعیت اولیه از راه‌حل‌ها است. سپس، کروموزوم‌ها به‌صورت درختی (ETs) نشان داده می شوند. میزان سازگاری هر عضو از جمعیت کروموزوم‌ها با تابع برازش تعیین می‌شود. در ادامه، باید تعداد ژن‌ها و کروموزوم‌ها برای اجرای مدل GEP تعیین شوند (۷). مراحل اصلی الگوریتم مدل برنامه‌ریزی بیان ژن در شکل‌های ۲ و ۳ نشان داده شده است.

فرایند گام به گام حل یک مسئله با استفاده از برنامه‌ریزی بیان ژن متشکل از پنج مرحله است (۱):

۱- انتخاب مجموعه ترمینال: که همان متغیرهای مستقل مسئله و متغیرهای حالت سامانه است. انتخاب تابع برازش در این مرحله صورت می‌گیرد که معمولاً از جذر میانگین مربعات خطا (Root mean square error, RMSE) استفاده می‌شود.

۲- انتخاب مجموعه توابع: که شامل عملگرهای ریاضی و توابع آزمون است. عملگرهای ریاضی شامل ۱۰ عملگر ضرب، تقسیم، جمع، تفریق، جذر، لگاریتم، مجذور، مکعب و ... هستند که سه عمل جمع، تفریق و ضرب بیشترین استفاده را دارند. در این پژوهش، از عملگرهای ضرب، جمع و تفریق استفاده شده است.

۳- انتخاب ساختار کروموزومی: ساختار کروموزومی از تعداد کروموزوم‌ها، اندازه سر (Head) و تعداد ژن‌ها تشکیل شده است.

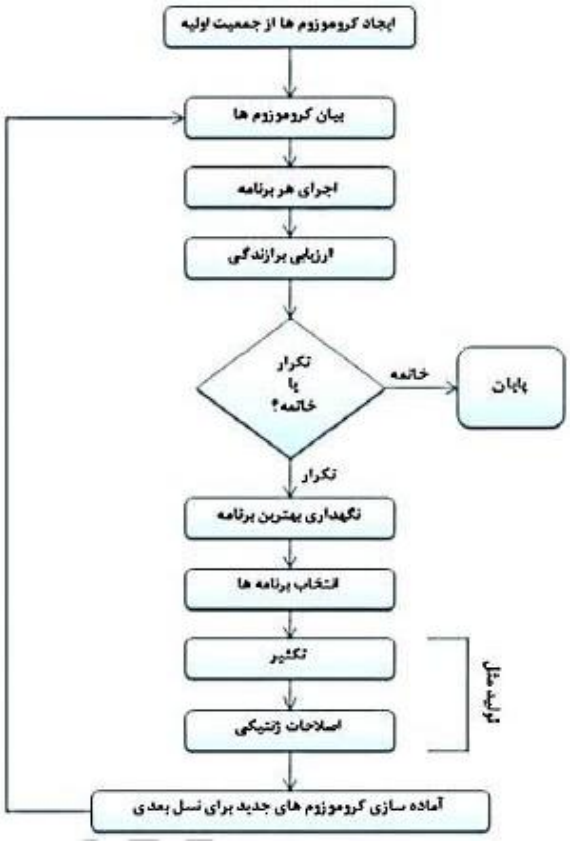
۴- انتخاب عملگر پیوند: شامل جمع (Addition)، تفریق (Subtraction)، ضرب (Multiplication) و تقسیم (Division) هستند.

۵- انتخاب عملگرهای ژنتیکی: شامل نرخ جهش، نرخ وارون سازی، نرخ ترکیب تک‌نقطه‌ای، نرخ ترکیب دونقطه‌ای، نرخ ترکیب ژن، نرخ ترانهش ژن، نرخ ترانهش درج متوالی و نرخ ترانهش درج ریشه.

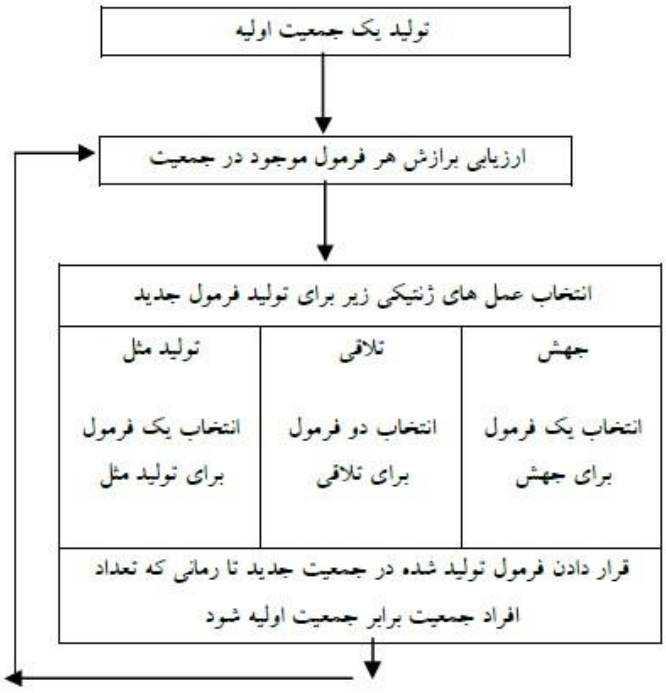
رگرسیون پارامتریک

رگرسیون خطی

روش‌های درون‌یابی در مهندسی آب قدمت زیادی داشته و برای حل و مدل‌سازی پارامترهای آن مورد استفاده قرار گرفته اند. این روش‌ها که عموماً مبتنی بر ارائه روابط خطی و غیرخطی بین ورودی‌های سیستم و پاسخ‌های آن هستند از تنوع فراوانی برخوردارند. رگرسیون خطی (Linear regression, LR) یک رویکرد خطی بین متغیر پاسخ با یک یا چند متغیر توصیفی است. اغلب برای کشف مدل رابطهٔ خطی بین متغیرها از رگرسیون خطی استفاده می‌شود. در این حالت، فرض بر این است که یک یا چندمتغیر توصیفی که مقدار آنها



شکل ۲. الگوریتم برنامه‌ریزی بیان ژن (۷)



شکل ۳. مروری بر شکل کلی اجرای برنامه‌ریزی بیان ژن (۱۸)

مستقل از بقیه متغیرها یا تحت کنترل محقق است، می‌تواند در پیش‌بینی متغیر پاسخ که مقدارش وابسته به متغیرهای توصیفی و تحت کنترل محقق نیست، مؤثر باشد. تفاوت رگرسیون و همبستگی این است که در همبستگی میزان و شدت رابطه دو یا چند متغیر مورد بررسی قرار می‌گیرد. اما در رگرسیون، پیش‌بینی یک یا چند متغیر بر اساس یک یا چند متغیر دیگر و بر پایه داده‌های گذشته انجام می‌شود (۲۳). لازم به‌ذکر است که تعداد مشاهدات (تعداد ردیف‌های) مرتبط با متغیرهای x و y باید با هم برابر باشند. پیش‌فرض رابطه بین x و y به‌صورت y= x+c است که در آن x به‌صورت بردار و c ثابت رگرسیون است.

رگرسیون غیرخطی

رگرسیون غیرخطی (NLR) روشی برای یافتن مدلی غیرخطی میان متغیر وابسته و مجموعه‌ای از متغیرهای مستقل است. بر خلاف روش LR که محاسبه مدل را محدود می‌کند، این نوع رگرسیون می‌تواند روابط مدل را به‌صورت اختیاری میان متغیرهای مستقل و غیرمستقل بررسی و اندازه‌گیری کند. مزایای رگرسیون غیرخطی عبارتند از: ۱) بیان فرایندهای فیزیکی، ۲) ارائه برآورد مناسب از پارامترهای مجهول در مدل با بهره‌گیری از مجموعه کوچکی از داده‌ها و ۳) توانایی استفاده از داده‌های کمی و کیفی (۲۳).

K- نزدیک‌ترین همسایگی

روش K نزدیک‌ترین همسایگی (K-NN) یکی از مهم‌ترین و توسعه‌یافته‌ترین رویکردهای ناپارامتریک است که در بسیاری از پژوهش‌های نوین برای تشخیص الگو و کلاسه‌بندی آماری به‌کار گرفته شده است (۱۹). اگر یک سری زمانی هیدرولوژیک معین نظیر:

$$D_N = \left\{ \{yi, xi\} \in R^1_{+} \times R^1_{+,}\ i = 1, \ldots, N \right\} \qquad (۶)$$

در دست باشد، یک تخمین‌گر رگرسیونی ناپارامتری، y^{NP} تابع هزینه (Cost function) زیر را روی بردار سری‌های زمانی D_N (داده‌های آموزشی) حداقل می‌کند:

$$c(y^{NP}) = \frac{\sum_{i=0}^{n} (y_i - y^{NP})^2 K(Xi, X0, b)}{N} \qquad (۷)$$

در این رابطه، متغیر خروجی با y_i و متغیرهای ورودی به‌صورت بردار Xi نشان داده شده‌اند و X0 مشخص کننده نقطه پرسش (Query point) بوده و از مجموعه داده‌های آزمون در فضای ورودی انتخاب می‌شود. متغیرهای ورودی را می‌توان بارش، جریان، جریان با تأخیرهای مختلف و یا هر متغیر دیگر، به طوری که ترکیبی از این متغیرها در یک یا چند ایستگاه باشد، درنظر گرفت. مقادیر متغیر خروجی با توجه به X0 محاسبه می شوند (۱۶).

مشابه D_N، مجموعه دیگری تحت عنوان داده‌های آزمون (D_M) درنظر گرفته می‌شود. به‌طوری که این داده‌ها هیچ عضو مشترکی با داده‌های بخش آموزشی (D_N) ندارند و نقطه پرسش نیز از بین این داده‌ها (مجموعه داده‌های آزمون) انتخاب می‌شود و متناظر با متغیر مستقل مربوط به بردار X است. همچنین، متغیر وابسته y در مجموعه داده‌های آزمون برای تست تخمین گر با توجه به اطلاعات داده‌های بخش آموزش (D_N) استفاده می‌شود.

تابع K نیز معرف وزن یا تابع کرنل است که مقدار آن با توجه به فاصله اقلیدسی هر نقطه در بخش آموزش از نقطه پرسش در بخش آزمون طبق رابطه ۸ به‌دست می‌آید:

$$K(X_i, X_0, b) = \begin{cases} 1 & \text{if} \quad X_i - X_0 \le b \\ 0 & \text{if} \quad X_i - X_0 > b \end{cases} \qquad (۸)$$

در این رابطه، $\|X_i - X_0\|$ نشان‌دهنده فاصله اقلیدسی و b شعاع همسایگی است. کمینه کردن رابطه ۸ با توجه به مقدار پارامتر y^{NP} کرنل تعریف شده در رابطه ۷ انجام می‌شود. برای تخمین y^{nn} از رابطه ۹ استفاده می‌شود:

$$y^{nn}_{Pred} = \frac{\sum_{i \in I_{nn}} y_i}{|I_{nn}|} \qquad (۹)$$

که در آن I_{nn} مجموعه‌ای است که عضوهای آن، آن تعداد از داده‌های مشاهداتی است که در داخل دایره‌ای به شعاع b از نقطه پرسش قرار دارند و $|I_{nn}|$ تعداد عضوهای مربوط به

جدول ۱. مقادیر پارامترهای آماری داده‌های روزانه جریان رودخانه کارون در ایستگاه ملاثانی (۹۶-۱۳۴۶)

نام دوره	تعداد داده آماری	میانگین (m^3/s)	حداکثر (m^3/s)	حداقل (m^3/s)	انحراف معیار (m^3/s)	ضریب تغییرات (٪)	چولگی (m^3/s)
آموزش	۱۴۹۰۰	۶۷۹/۱۸	۶۴۶۲	۴۲/۶	۵۸۷/۸۲	۰/۸۶	۲/۹۵
صحت‌سنجی	۳۷۲۳	۴۰۳/۷۸	۳۴۸۵/۸۳	۸۶	۳۷۲/۹۳	۰/۹۲	۳/۶۵

مجموعه I_{nn} است. یعنی:

$$\left|I_{nn}\right|\{i:X_i - X_0 \le b\} \quad (۱۰)$$

کارایی عملکرد این روش بستگی به انتخاب پارامترهای b (شعاع همسایگی) و I (تعداد تأخیرها در پارامترهای ورودی) دارد (۱۶).

معیارهای ارزیابی الگوها

نمایه‌های ضریب تبیین (Determination coefficient, R^2) جذر میانگین مربعات خطا (RMSE) و متوسط خطای مطلق (Mean absolute error, MAE) برای ارزیابی الگوهای مورد مطالعه در این تحقیق استفاده شده‌اند (روابط ۱۱ تا ۱۳). درصورتی که مقدار ضریب تبیین زیاد و ضرایب خطا کم باشد می‌توان نتایج دقیق‌تر و قابل اعتمادتری به‌دست آورد:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Q_O - Q_P)^2}{\sum_{i=1}^{n}(Q_O - \bar{Q})^2} \quad (۱۱)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Q_O - Q_P)^2}{n}} \quad (۱۲)$$

$$MAE = \frac{\sum_{i=1}^{n}|Q_O - Q_P|}{n} \quad (۱۳)$$

در روابط بالا Q_O مقدار دبی مشاهداتی رودخانه در گام iام، Q_p مقدار دبی پیش‌بینی شده در همان زمان، n تعداد داده‌ها و $\bar{Q}$ میانگین مقادیر دبی مشاهداتی است.

نتایج و بحث

در این مطالعه، از داده‌های میانگین دبی روزانه ۵۱ ساله ایستگاه هیدرومتری ملاثانی در سال‌های ۹۶-۱۳۴۶ استفاده شده است. از مجموع ۱۸۶۲۳ داده متوسط دبی روزانه رودخانه کارون در محل ایستگاه ملاثانی، تعداد ۱۴۹۰۰ داده به‌منظور آموزش و تعداد ۳۷۲۳ داده باقیمانده برای صحت‌سنجی مدل استفاده شد. به‌طور کلی ۸۰ درصد داده‌ها برای آموزش و ۲۰ درصد برای تست درنظر گرفته شده است. پارامترهای آماری دبی روزانه ایستگاه ملاثانی در جدول ۱ ارائه شده‌اند.

مطابق نتایج جدول ۱، اگر چه حداکثر دبی در دوره آموزش بیشتر از دوره آزمون است، ولی حداقل دبی در دوره آموزش کمتر از دوره آزمون بوده است. این موضوع برای برون‌یابی داده‌های مورد آزمون به‌منظور ارزیابی بهتر مدل‌های مورد استفاده در این تحقیق دارای اهمیت است (۱۳).

نتیجه بررسی روند داده‌های سری زمانی

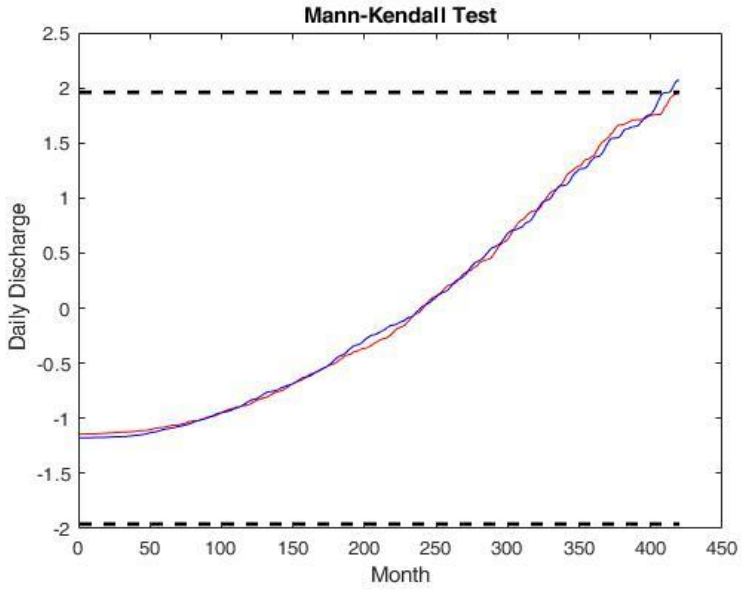
نتیجه بررسی روند داده‌های میانگین دبی روزانه ایستگاه هیدرومتری ملاثانی با استفاده از آزمون من- کندال نشان می‌دهد که داده‌های استفاده شده در این تحقیق در سطح اعتماد ۹۵ درصد فاقد روند بوده و تصادفی هستند (شکل ۴).

نتیجه آزمون سیر تناوبی

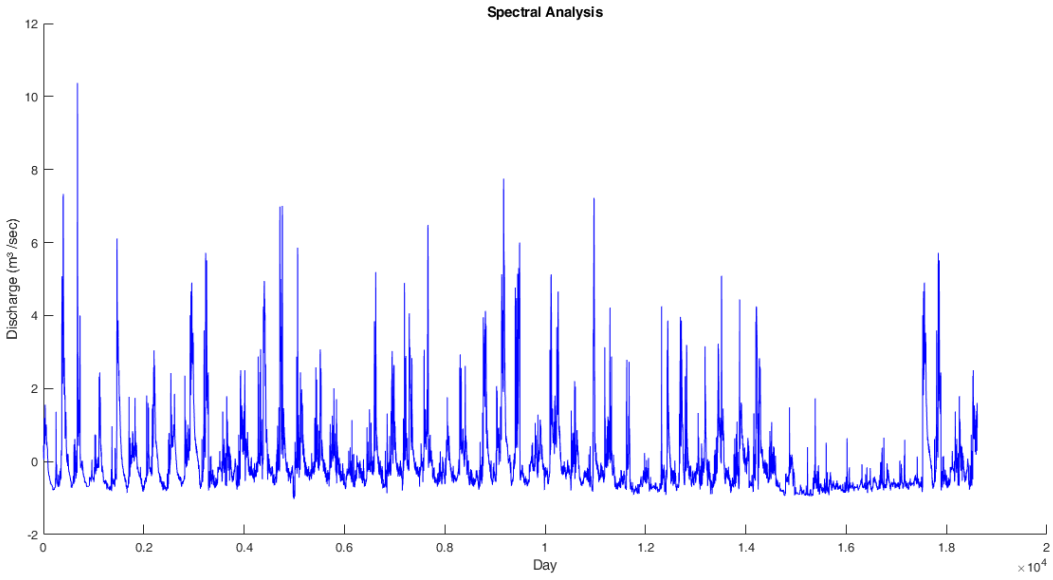
همان‌گونه که شکل ۵ نشان می‌دهد، سری زمانی مورد استفاده در این تحقیق فاقد سیر تناوبی بوده و قابلیت لازم برای استفاده در مدل‌سازی را دارد.

نتیجه آزمون کفایت داده‌ها

مقدار ضریب هرست برای داده‌های روزانه ایستگاه هیدرومتری ملاثانی که در این تحقیق استفاده شده برابر ۰/۸۴۳ است که نشان‌دهنده مناسب بودن حافظه بلندمدت طول دوره آماری



شکل ۴. نتایج آزمون من-کندال داده‌های میانگین دبی روزانه ایستگاه ملاثانی

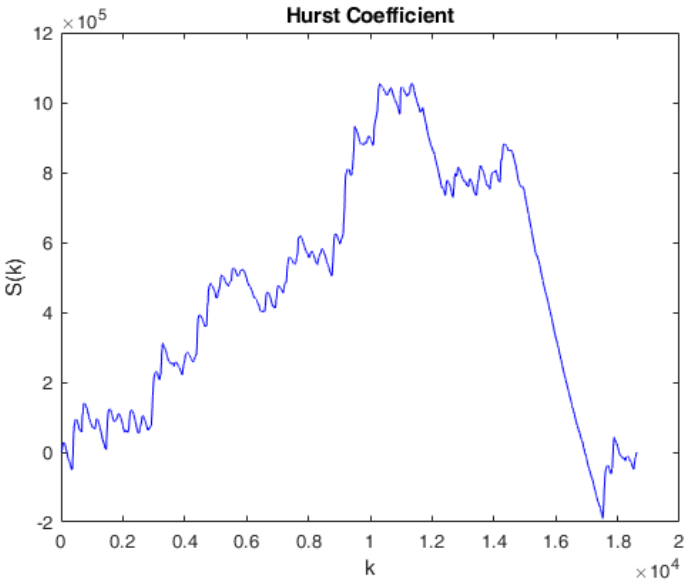


شکل ۵. سیر تناوبی داده‌های میانگین دبی روزانه در ایستگاه هیدرومتری ملاثانی

سری زمانی است (شکل ۶).

نتایج مدل‌سازی برنامه‌ریزی بیان ژن

در مدل برنامه‌ریزی بیان ژن، انتخاب جمعیت اولیه که همان الگوهای ورودی است از اهمیت بالایی برخوردار است. با توجه به اینکه در این مطالعه توالی دبی روزهای قبل در پیش‌بینی دبی امروز مد نظر بوده است فقط از داده‌های دبی جریان با توالی برگشتی تا ۶ روز قبل به‌عنوان داده‌های آموزشی به‌صورت ترکیب‌های مختلف استفاده شده است. همچنین، عملکرد این مدل بر اساس روش سعی و خطا، تعداد ۳۰ کروموزم، ۸ سر و ۳ ژن به‌عنوان بهترین ترکیب و دارای کوتاه ترین زمان محاسبه تعیین شد که در این مطالعه، عملگر



شکل ۶. نتایج آزمون هرست داده‌های میانگین دبی روزانه در ایستگاه ملاثانی

جدول ۲. تحلیل آماری نتایج الگوهای مدل برنامه‌ریزی بیان ژن برای پیش‌بینی میانگین دبی روزانه در ایستگاه ملاثانی

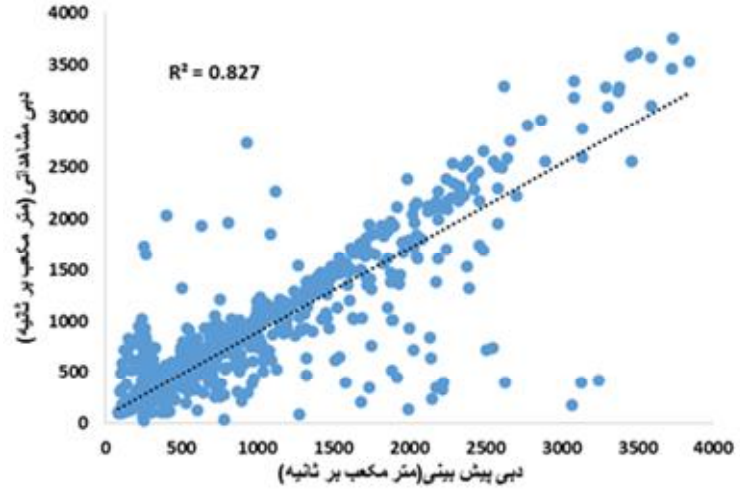
شماره الگو	الگوی ورودی روزانه	مرحله آموزش			مرحله صحت‌سنجی		
		R^2	MAE	RMSE	R^2	MAE	RMSE
۱	$Q(t)=f\{Q(t-1)\}$	۰/۷۸۶	۴۷/۰۴	۱۲۲/۰۰	۰/۸۰۱	۴۲/۱۶	۱۱۹/۰۱
۲	$Q(t)=f\{Q(t-1),Q(t-2)\}$	۰/۷۹۱	۴۰/۳۷	۱۱۰/۸۸	۰/۸۰۳	۳۸/۴۵	۸۵/۱۱
۳	$Q(t)=f\{Q(t-1),Q(t-2),Q(t-3)\}$	۰/۷۹۱	۴۰/۰۹	۱۱۲/۵۳	۰/۸۱۰	۳۵/۳۵	۶۷/۹۲
۴	$Q(t)=f\{Q(t-1),Q(t-2),Q(t-3),Q(t-4)\}$	۰/۸۰۱	۳۹/۶۴	۹۵/۷۱	۰/۸۱۴	۳۳/۶۷	۶۴/۱۰
۵	$Q(t)=f\{Q(t-1),Q(t-2),Q(t-3),Q(t-4),Q(t-5)\}$	۰/۸۱۴	۳۷/۲۲	۷۴/۳۲	۰/۸۲۷	۲۶/۶۴	۵۹/۴۵
۶	$Q(t)=f\{Q(t-1),Q(t-2),Q(t-3),Q(t-4),Q(t-5),Q(t-6)\}$	۰/۷۱۱	۶۴/۸۲	۱۴۹/۳۷	۰/۷۹۸	۵۹/۴۱	۱۸۸/۰۲

جمع بهترین عملکرد را ارائه داد. در این مطالعه، معیار توقف بر اساس میزان جمعیت تولید شده (Generation Number) برابر ۱۰۰۰ تعیین شده است. نرم‌افزار مورد استفاده در این تحقیق، GeneXproTools 5.0 بوده است.

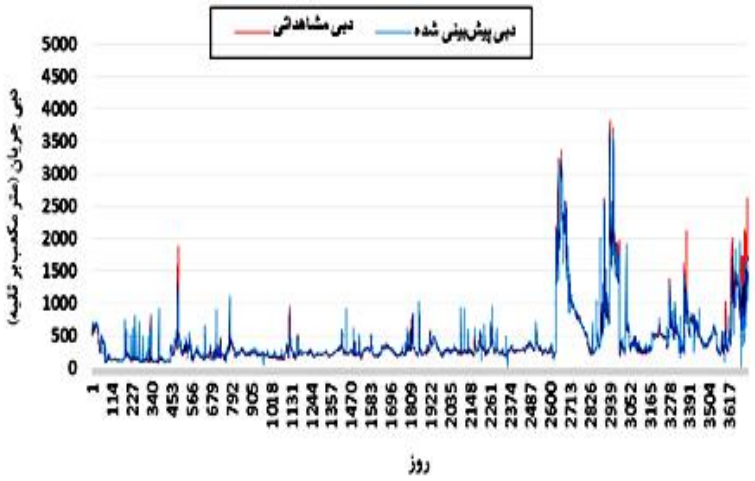
نتایج ارائه شده در جدول ۲ در مراحل آموزش و صحت‌سنجی نشان می‌دهد که الگوی شماره ۵ با R^2 =۰/۸۲۷، RMSE = ۵۹/۴۵ و MAE= ۲۶/۶۴ دارای بهترین عملکرد در مرحله صحت‌سنجی بوده که به‌عنوان بهترین الگو برای مدل برنامه‌ریزی بیان ژن انتخاب شده است. شکل ۷ نمودار پراکندگی و شکل ۸ نمودار مقادیر مشاهداتی و پیش‌بینی شده حاصل از بهترین مدل برنامه‌ریزی بیان ژن برای الگوی ۵ در مرحله صحت‌سنجی را نشان می‌دهد. همان‌گونه که شکل ۸ نشان می‌دهد، دقت پیش‌بینی مدل برنامه‌ریزی بیان ژن در پیش‌بینی دبی‌های کمتر جریان، بهتر از دبی‌های بزرگ‌تر جریان بوده است.

نتایج مدل‌سازی رگرسیون خطی و غیرخطی

مطابق جدول‌های ۳ و ۴، به‌ازای الگوهای ورودی مختلف، بهترین عملکرد هر دو روش رگرسیون خطی و غیرخطی در پیش‌بینی جریان روزانه ایستگاه ملاثانی در الگوی شماره ۵ بوده است. بر



شکل ۷. نمودار پراکندگی مقادیر مشاهداتی و پیش‌بینی شده دبی روزانه ایستگاه ملاثانی با استفاده از مدل برنامه‌ریزی بیان ژن در مرحله صحت‌سنجی



شکل ۸. مقادیر مشاهداتی و پیش‌بینی شده دبی روزانه ایستگاه ملاثانی با استفاده از مدل برنامه‌ریزی بیان ژن در مرحله صحت‌سنجی (رنگی در نسخه الکترونیکی)

جدول ۳. تحلیل آماری رگرسیون خطی برای الگوهای مختلف ورودی میانگین دبی روزانه در ایستگاه ملاثانی

شماره الگو	الگوی ورودی روزانه	مرحله آموزش			مرحله صحت‌سنجی		
		R^2	MAE	RMSE	R^2	MAE	RMSE
۱	$Q(t)=f\{Q(t-1)\}$	۰/۷۲۰	۷۷/۴۹	۶۸/۳۲	۰/۷۳۴	۶۲/۱۸	۱۳۳/۳۵
۲	$Q(t)=f\{Q(t-1),Q(t-2)\}$	۰/۷۳۵	۷۱/۰۵	۶۲/۱۱	۰/۷۳۷	۶۰/۱۹	۱۲۹/۴۶
۳	$Q(t)=f\{Q(t-1),Q(t-2),Q(t-3)\}$	۰/۷۴۳	۶۹/۹۷	۵۸/۲۲	۰/۷۶۹	۴۸/۸۳	۱۰۹/۱۴
۴	$Q(t)=f\{Q(t-1),Q(t-2),Q(t-3),Q(t-4)\}$	۰/۷۶۹	۴۸/۲۵	۴۸/۷۷	۰/۷۸۹	۴۵/۱۱	۱۰۲/۱۷
۵	$Q(t)=f\{Q(t-1),Q(t-2),Q(t-3),Q(t-4),Q(t-5)\}$	۰/۷۹۱	۴۱/۳۷	۳۹/۷۹	۰/۸۰۰	۳۶/۹۴	۹۱/۱۰
۶	$Q(t)=f\{Q(t-1),Q(t-2),Q(t-3),Q(t-4),Q(t-5),Q(t-6)\}$	۰/۷۲۵	۶۵/۹۷	۷۲/۰۷	۰/۷۳۱	۶۴/۲۶	۱۳۵/۰۰

جدول ۴. تحلیل آماری رگرسیون غیرخطی برای الگوهای مختلف ورودی میانگین دبی روزانه در ایستگاه ملاثانی

شماره الگو	الگوی ورودی روزانه	مرحله آموزش			مرحله صحت‌سنجی		
		R^2	MAE	RMSE	R^2	MAE	RMSE
۱	$Q(t) = f\{Q(t-1)\}$	۰/۶۱۱	۹۹/۳۸	۲۷۸/۱۴	۰/۶۲۱	۹۰/۹۳	۲۶۴/۲۶
۲	$Q(t) = f\{Q(t-1), Q(t-2)\}$	۰/۶۴۸	۸۷/۵۴	۲۳۲/۷۱	۰/۶۴۹	۸۷/۲۵	۲۲۸/۹۳
۳	$Q(t) = f\{Q(t-1), Q(t-2), Q(t-3)\}$	۰/۶۵۹	۸۳/۶۸	۲۱۰/۱۵	۰/۶۸۴	۷۶/۳۹	۱۹۵/۲۲
۴	$Q(t) = f\{Q(t-1), Q(t-2), Q(t-3), Q(t-4)\}$	۰/۶۷۴	۷۷/۰۰	۱۹۷/۲۰	۰/۶۹۲	۷۰/۷۱	۱۸۳/۴۱
۵	$Q(t) = f\{Q(t-1), Q(t-2), Q(t-3), Q(t-4), Q(t-5)\}$	۰/۶۷۹	۷۱/۳۷	۱۸۴/۲۵	۰/۶۹۵	۷۰/۲۸	۱۸۱/۶۹
۶	$Q(t) = f\{Q(t-1), Q(t-2), Q(t-3), Q(t-4), Q(t-5), Q(t-6)\}$	۰/۶۴۳	۸۸/۰۲	۹۴/۲۴	۰/۵۱۷	۱۲۴/۱۴	۵۲۳/۱۰

این اساس، در مرحله صحت‌سنجی، روش رگرسیون خطی با R^2 = ۰/۸۰۰، RMSE = ۹۱/۱۰ و MAE= ۳۶/۹۴ عملکرد بهتری نسبت به رگرسیون غیرخطی داشته است.

شکل‌های ۹ تا ۱۲ نشان می‌دهند که هر دو روش رگرسیونی، میزان جریان در ایستگاه ملاثانی را در اغلب موارد بیش از مقدار واقعی آن پیش‌بینی کرده‌اند. که ناشی از تشخیص نامناسب پیش‌بینی تغییرات جریان توسط روش‌های رگرسیونی است.

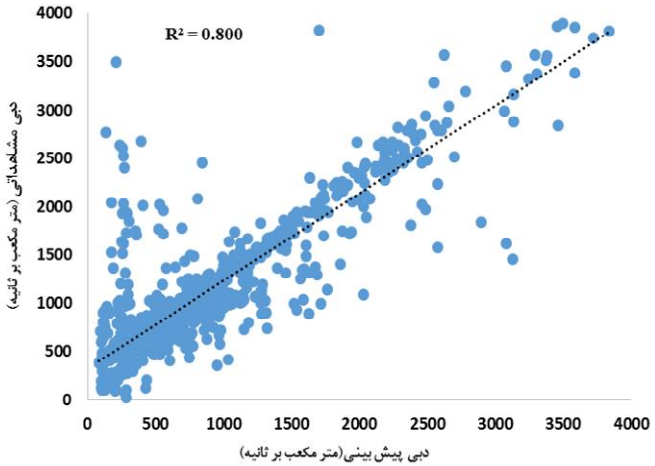
نتایج مدل‌سازی K-NN

در این روش، پیش‌بینی میزان جریان، با توجه به تعداد تأخیرها و همچنین شعاع همسایگی انجام می‌شود. در شکل ۱۳ تغییرات و پراکندگی داده‌های مشاهداتی و پیش‌بینی شده جریان توسط مدل نزدیک‌ترین همسایگی به‌ازای بهینه‌ترین ساختار مدل (الگوی ۵) نشان داده شده است. نتایج ارائه شده در جدول ۵ نشان می‌دهد که بهترین الگوی پیش‌بینی جریان دارای R^2 = ۰/۸۲۳، RMSE = ۴۹/۷۴ و MAE= ۳۷/۲۹ در مرحله صحت‌سنجی است.

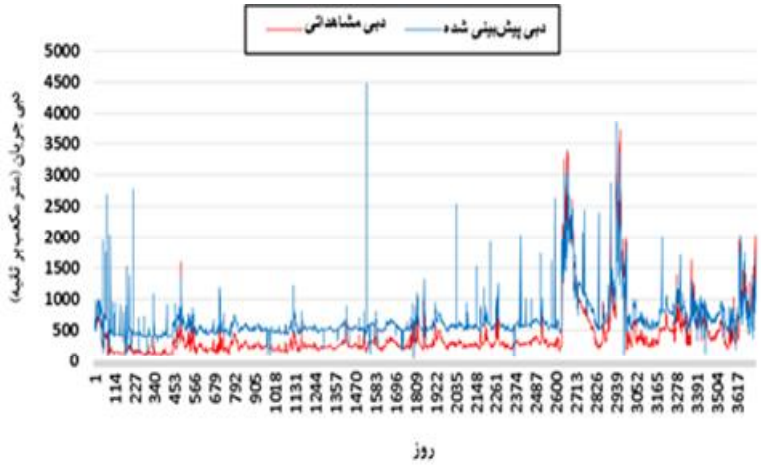
با توجه به شکل ۱۴ می‌توان دریافت که عملکرد مدل در پیش‌بینی مقادیر حداقل جریان به‌مراتب بهتر از پیش‌بینی مقادیر حداکثر آن بوده که این مسئله می‌تواند ناشی از وجود مقادیر دبی متناظر با مقادیر حداقل جریان در همسایگی نقطه مورد پیش‌بینی باشد که در عملکرد مدل تأثیر زیادی داشته است.

بررسی نتایج جدول‌های ۲ تا ۵ نشان می‌دهد که دقت مدل برنامه‌ریزی بیان ژن در پیش‌بینی جریان روزانه ایستگاه ملاثانی نسبت به سایر روش‌های مورد استفاده در این تحقیق مطلوب‌تر بوده است.

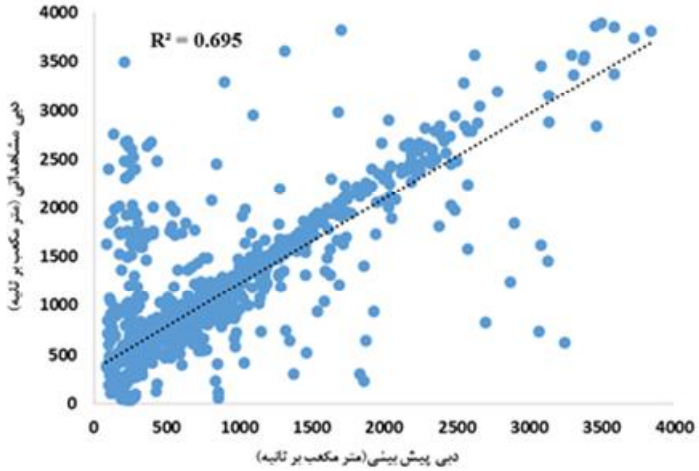
به‌منظور بررسی اثر کنترل‌کنندگی سدهای بالادست ایستگاه ملاثانی (جدول ۶) بر نتایج پیش‌بینی میانگین دبی روزانه رودخانه کارون، داده‌های میانگین روزانه ایستگاه هیدرومتری ملاثانی با روش‌های مورد استفاده در این تحقیق ارزیابی شده‌اند. برای مثال، تحلیل آماری الگوهای مدل برنامه‌ریزی بیان ژن بعد از احداث سد شهید عباسپور برای پیش‌بینی دبی روزانه ایستگاه ملاثانی (۹۶-۱۳۵۶) در جدول ۷ و تحلیل آماری الگوهای مدل برنامه‌ریزی بیان ژن بعد از احداث سد کارون ۳ برای پیش‌بینی دبی روزانه ایستگاه ملاثانی (۹۶-۱۳۸۴) در جدول ۸ نشان داده شده است. نتایج نشان داد که پیش‌بینی میانگین دبی جریان روزانه ایستگاه ملاثانی با استفاده از برنامه‌ریزی بیان ژن دارای دقت مناسبی بوده و احداث سدها در بالادست ایستگاه مورد مطالعه با توجه به حجم و شیوه بهره‌برداری



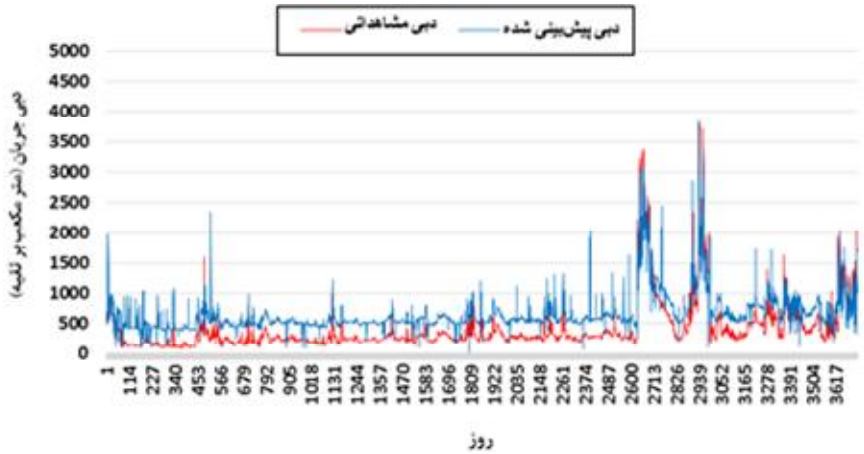
شکل ۹. نمودار پراکندگی مقادیر مشاهداتی و پیش‌بینی شده دبی روزانه ایستگاه ملاثانی با استفاده از رگرسیون خطی در مرحله صحت‌سنجی



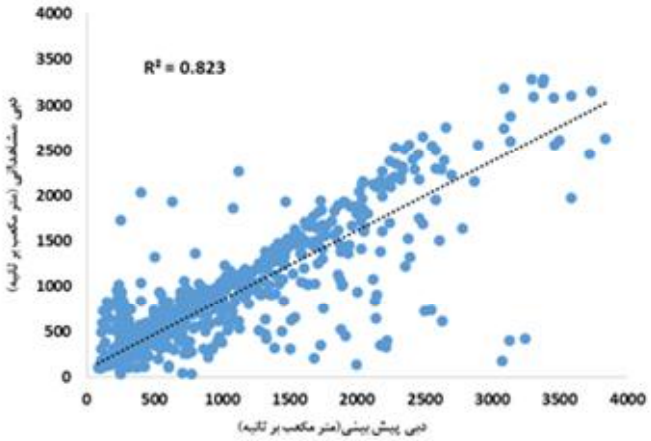
شکل ۱۰. مقادیر مشاهداتی و پیش‌بینی شده دبی روزانه ایستگاه ملاثانی با استفاده از رگرسیون خطی در مرحله صحت‌سنجی (رنگی در نسخه الکترونیکی)



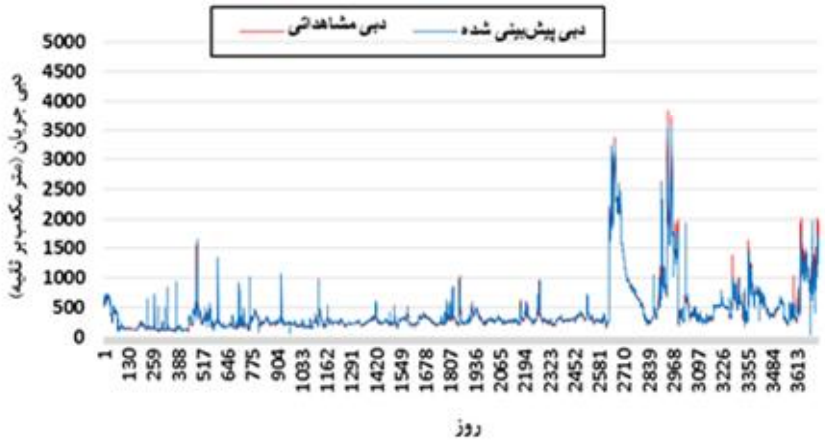
شکل ۱۱. نمودار پراکندگی مقادیر مشاهداتی و پیش‌بینی شده دبی روزانه ایستگاه ملاثانی با استفاده از رگرسیون غیرخطی در مرحله صحت‌سنجی



شکل ۱۲. مقادیر مشاهداتی و پیش‌بینی شده دبی روزانه ایستگاه ملاثانی با استفاده از رگرسیون غیرخطی در مرحله صحت‌سنجی (رنگی در نسخه الکترونیکی)



شکل ۱۳. نمودار پراکندگی مقادیر مشاهداتی و پیش‌بینی شده دبی روزانه ایستگاه ملاثانی با استفاده از الگوریتم K –نزدیک‌ترین همسایگی در مرحله صحت‌سنجی



شکل ۱۴. مقادیر مشاهداتی و پیش‌بینی شده دبی روزانه ایستگاه ملاثانی با استفاده از الگویتم K- نزدیک‌ترین همسایگی در مرحله صحت‌سنجی (رنگی در نسخه الکترونیکی)

جدول ۵. تحلیل آماری الگوریتم K-NN برای الگوهای مختلف ورودی میانگین دبی روزانه در ایستگاه ملاثانی

شماره الگو	الگوی ورودی روزانه	مرحله آموزش			مرحله صحت سنجی		
		R^2	MAE	RMSE	R^2	MAE	RMSE
۱	$Q(t) = f\{Q(t-1)\}$	۰/۷۰۷	۸۸/۷۶	۲۰۵/۴۱	۰/۷۵۳	۸۰/۵۲	۱۴۶/۰۴
۲	$Q(t) = f\{Q(t-1), Q(t-2)\}$	۰/۷۳۲	۸۲/۴۰	۱۶۱/۲۸	۰/۷۵۶	۷۸/۰۰	۱۴۲/۷۰
۳	$Q(t) = f\{Q(t-1), Q(t-2), Q(t-3)\}$	۰/۷۵۲	۶۱/۹۵	۱۰۰/۱۱	۰/۷۵۸	۷۵/۷۳	۱۳۳/۵۵
۴	$Q(t) = f\{Q(t-1), Q(t-2), Q(t-3), Q(t-4)\}$	۰/۷۶۳	۶۹/۴۲	۱۲۳/۰۳	۰/۷۸۸	۶۶/۳۲	۱۰۵/۱۴
۵	$Q(t) = f\{Q(t-1), Q(t-2), Q(t-3), Q(t-4), Q(t-5)\}$	۰/۸۱۴	۶۱/۱۹	۹۸/۴۱	۰/۸۲۳	۳۷/۲۹	۴۹/۷۴
۶	$Q(t) = f\{Q(t-1), Q(t-2), Q(t-3), Q(t-4), Q(t-5), Q(t-6)\}$	۰/۷۰۳	۶۱/۳۹	۱۹۳/۶۲	۰/۷۳۶	۸۱/۳۸	۱۵۲/۸۹

جدول ۶. نام سدهای مخزنی ساخته شده در بالادست ایستگاه هیدرومتری ملاثانی و زمان بهره‌برداری از آنها

ردیف	نام سد	زمان بهره‌برداری	حجم مخزن (میلیارد متر مکعب)	محل قرارگیری سد
۱	دز	اسفندماه ۱۳۴۱	۳/۳۰	شاخه دز رودخانه کارون بزرگ
۲	شهید عباسپور (کارون ۱)	آذرماه ۱۳۵۵	۳/۱۴	شاخه کارون رودخانه کارون بزرگ
۳	کارون ۳	اسفندماه ۱۳۸۳	۳/۰۰	شاخه کارون رودخانه کارون بزرگ
۴	مسجد سلیمان (گدار لندر)	فروردین‌ماه ۱۳۸۰	۲/۶۵	شاخه کارون رودخانه کارون بزرگ
۵	گتوند علیا	مردادماه ۱۳۹۰	۴/۵۰	شاخه کارون رودخانه کارون بزرگ

جدول ۷. تحلیل آماری الگوهای مدل برنامه‌ریزی بیان ژن بعد از احداث سد شهید عباسپور برای پیش‌بینی میانگین دبی روزانه ایستگاه ملاثانی (۹۶-۱۳۵۶)

شماره الگو	الگوی ورودی روزانه	مرحله آموزش			مرحله صحت‌سنجی		
		R^2	MAE	RMSE	R^2	MAE	RMSE
۱	$Q(t) = f\{Q(t-1)\}$	۰/۸۱۶	۲۵/۱۲	۷۵/۱۶	۰/۸۱۵	۲۷/۰۰	۶۹/۹۱
۲	$Q(t) = f\{Q(t-1), Q(t-2)\}$	۰/۸۱۹	۲۲/۴۹	۶۹/۴۲	۰/۸۲۲	۲۰/۱۷	۵۳/۱۱
۳	$Q(t) = f\{Q(t-1), Q(t-2), Q(t-3)\}$	۰/۸۳۰	۱۹/۲۴	۴۳/۶۹	۰/۸۴۱	۱۴/۰۸	۳۹/۰۰
۴	$Q(t) = f\{Q(t-1), Q(t-2), Q(t-3), Q(t-4)\}$	۰/۸۳۵	۱۹/۰۸	۴۱/۱۱	۰/۸۵۶	۱۴/۰۱	۳۰/۹۷
۵	$Q(t) = f\{Q(t-1), Q(t-2), Q(t-3), Q(t-4), Q(t-5)\}$	۰/۸۲۴	۲۱/۶۷	۴۹/۶۳	۰/۸۳۷	۱۹/۲۵	۴۳/۷۵
۶	$Q(t) = f\{Q(t-1), Q(t-2), Q(t-3), Q(t-4), Q(t-5), Q(t-6)\}$	۰/۷۹۴	۵۶/۱۷	۹۳/۵۴	۰/۷۹۹	۴۶/۰۳	۸۵/۶۲

جدول ۸. تحلیل آماری الگوهای مدل برنامه‌ریزی بیان ژن بعد از احداث سد کارون ۳ برای پیش‌بینی میانگین دبی روزانه ایستگاه ملاثانی (۹۶-۱۳۸۴)

شماره الگو	الگوی ورودی روزانه	مرحله آموزش			مرحله صحت‌سنجی		
		R^2	MAE	RMSE	R^2	MAE	RMSE
۱	$Q(t) = f\{Q(t-1)\}$	۰/۷۴۲	۶۱/۸۸	۱۳۲/۶۴	۰/۷۶۳	۴۶/۲۱	۷۹/۲۴
۲	$Q(t) = f\{Q(t-1), Q(t-2)\}$	۰/۷۵۶	۵۷/۰۵	۱۱۹/۰۹	۰/۷۷۹	۴۷/۳۰	۶۲/۹۷
۳	$Q(t) = f\{Q(t-1), Q(t-2), Q(t-3)\}$	۰/۷۶۱	۵۱/۰۰	۸۵/۳۱	۰/۷۸۴	۳۵/۱۱	۵۴/۲۱
۴	$Q(t) = f\{Q(t-1), Q(t-2), Q(t-3), Q(t-4)\}$	۰/۷۶۹	۴۲/۸۶	۶۳/۰۰	۰/۷۸۷	۳۳/۷۵	۵۴/۰۸
۵	$Q(t) = f\{Q(t-1), Q(t-2), Q(t-3), Q(t-4), Q(t-5)\}$	۰/۷۶۵	۴۹/۱۹	۷۶/۳۷	۰/۷۶۹	۴۱/۵۹	۶۸/۱۷
۶	$Q(t) = f\{Q(t-1), Q(t-2), Q(t-3), Q(t-4), Q(t-5), Q(t-6)\}$	۰/۷۴۴	۵۳/۳۹	۷۹/۹۲	۰/۷۵۱	۵۱/۰۰	۷۳/۳۳

جدول ۹. پیش‌بینی مقادیر حداکثر جریان ایستگاه ملاثانی با مدل‌های مورد استفاده و درصد خطای نسبی متناظر با آنها

ردیف	حداکثر جریان (m^3/s)	تاریخ رویداد	برنامه‌ریزی بیان ژن (m^3/s)	رگرسیون خطی (m^3/s)	رگرسیون غیرخطی (m^3/s)	الگوریتم K-NN (m^3/s)	خطای نسبی برنامه‌ریزی بیان ژن	خطای نسبی رگرسیون خطی	خطای نسبی رگرسیون غیرخطی	خطای نسبی الگوریتم K-NN
۱	۳۰۰۵	۱۳۹۲/۱۱/۲۸	۲۹۳۵	۲۴۳۲	۲۱۳۰	۲۸۸۱	-۲/۳۳	-۱۹/۰۷	-۲۹/۱۲	-۴/۱۳
۲	۱۷۰۲	۱۳۹۳/۰۱/۱۷	۱۵۹۶	۲۳۸۹	۲۱۴۹	۱۵۰۴	-۶/۲۳	۴۰/۳۶	۲۶/۲۶	-۱۱/۶۳
۳	۳۲۴۷	۱۳۹۴/۰۱/۱۹	۳۰۳۷	۲۶۳۷	۲۲۱۹	۳۰۲۸	-۶/۴۷	-۱۸/۷۹	-۳۱/۶۶	-۶/۷۴
۴	۳۱۳۹	۱۳۹۴/۱۱/۴	۳۲۴۵	۳۷۱۴	۳۴۵۰	۳۱۵۷	۳/۳۷	۱۸/۳۲	۹/۹۱	۵/۷۳
۵	۱۳۹۲	۱۳۹۵/۱۰/۱۰	۱۳۸۱	۱۲۹۵	۱۱۰۴	۱۳۷۶	-۰/۷۹	-۶/۹۷	-۲۰/۶۹	-۱/۱۵
۶	۱۶۲۹	۱۳۹۶/۰۱/۰۱	۱۵۷۳	۱۵۰۶	۱۳۱۹	۱۶۱۳	-۳/۴۴	-۷/۵۵	-۱۹/۰۳	-۰/۹۸
۷	۲۰۳۰	۱۳۹۶/۱۰/۰۲	۱۹۶۰	۱۷۱۹	۱۷۰۴	۱۹۰۳	-۳/۴۵	-۱۵/۳۲	-۱۶/۰۶	-۶/۲۶
	متوسط قدر مطلق خطا						۳/۷۳	۱۸/۰۵	۲۱/۸۲	۵/۲۳

از آنها، تأثیر قابل ملاحظه‌ای بر دقت پیش‌بینی مطلوب میزان جریان رودخانه نداشته است و فقط اثر احداث سدها و بهره‌برداری از آنها بر نوع الگوی پیش‌بینی جریان (گام تأخیر روزانه) مؤثر بوده است. لذا، از ارائه سایر نتایج این بخش به‌دلیل افزایش قابل توجه حجم مطالب و تعداد جدول‌های این تحقیق خودداری شده است.

بررسی دقت مدل‌ها در پیش‌بینی حداکثر جریان

با توجه به اهمیت مقادیر حداکثر جریان برای به‌کارگیری در سیستم‌های هشدار سیل، طراحی سازه‌های آبی و مطالعات منابع آب، در این بخش از تحقیق به بررسی دقت مدل‌های مورد استفاده در پیش‌بینی مقادیر حداکثر جریان در ایستگاه ملاثانی پرداخته شده است. در جدول ۹، مقادیر مشاهداتی

حداکثر جریان و مقادیر پیش‌بینی شده با مدل‌های مختلف و دقت پیش‌بینی هر مدل ارائه شده است. قابل ذکر است که داده‌های انتخابی از مجموعه داده‌های مرحله صحت‌سنجی مدل‌ها بوده‌اند و نتایج ارائه شده برای مدل‌های مختلف به‌ازای بهینه‌ترین الگو (الگوی شماره ۵) هستند. میزان خطای نسبی نیز از رابطه ۱۴ محاسبه شده است:

$$RE = \frac{Q_P - Q_O}{Q_O} \qquad (۱۴)$$

که Q_O مقدار دبی مشاهداتی و Q_p مقدار دبی پیش‌بینی شده است.

تغییرات میزان متوسط قدر مطلق خطا برای مقادیر حداکثر جریان با استفاده از مدل‌های مختلف در مقایسه با مقادیر مشاهداتی در جدول ۹ نشان داده شده است. از این نتایج می‌توان دریافت که عموماً پیش‌بینی جریان حداکثر توسط مدل‌های مختلف کمتر از مقادیر مشاهداتی بوده که ممکن است این امر ناشی از ماهیت تصادفی مقادیر حداکثر جریان باشد.

نتایج جدول ۹ نشان می‌دهد که گرچه مدل برنامه‌ریزی بیان ژن و الگوریتم K- نزدیک‌ترین همسایگی در رابطه با پیش‌بینی روند تغییرات جریان در ایستگاه مورد مطالعه دقت تقریباً یکسانی داشته‌اند، لیکن در مورد پیش‌بینی دبی حداکثر جریان در ایستگاه ملاثانی، دقت مدل برنامه‌ریزی بیان ژن بهتر از الگوریتم K- نزدیک‌ترین همسایگی بوده است. بنابراین، مدل برنامه‌ریزی بیان ژن برای پیش‌بینی آبدهی روزانه ایستگاه ملاثانی پیشنهاد می‌شود.

اهمیت پژوهش حاضر در مقایسه با پژوهش‌های قبلی انجام گرفته در خصوص پیش‌بینی آبدهی انجام شده روی رودخانه کارون، از جمله پژوهش زمانی و همکاران (۲۵)، این است که مطالعات گذشته پیش‌بینی جریان، در ایستگاه‌های بالادست مخازن موجود روی این رودخانه و بیشتر در سرشاخه‌های آن انجام گرفته است. درحالی که این پژوهش روی ایستگاه هیدرومتری ملاثانی متمرکز بوده که در پایین‌دست پنج مخزن بزرگ و نیز دو شاخه مجزای رودخانه کارون بزرگ (رودخانه‌های دز و کارون) انجام شده که پیش‌بینی آبدهی روزانه آن با مدل‌های شبیه‌سازی هیدرولوژیک بارش- رواناب، با توجه به متفاوت بودن شرایط هیدرولوژیک و مورفولوژیک هر یک از حوضه‌ها و رودخانه‌های بالادست (دز و کارون)، یا عملاً امکان‌پذیر نبوده و یا نیاز به دستیابی به اطلاعات فراوان و صرف وقت زیاد داشته که درنهایت ممکن است خروجی‌های مدل نیز فاقد دقت لازم باشند. لیکن، درصورت استفاده از روش‌های داده‌مبنا در این تحقیق، می‌توان با صرف وقت اندک و داشتن دقت مطلوب، آبدهی روزانه ایستگاه ملاثانی را پیش‌بینی کرد.

نتیجه‌گیری

در این مطالعه، به بررسی عملکرد مدل برنامه‌ریزی بیان ژن (GEP)، الگوریتم K-NN و رگرسیون‌های خطی و غیرخطی در پیش‌بینی جریان روزانه رودخانه کارون در جنوب غرب ایران پرداخته شد. در این راستا، از داده‌های متوسط آبدهی روزانه ایستگاه ملاثانی واقع بر رودخانه کارون در فاصله سال‌های ۱۳۴۶ تا ۱۳۹۶ استفاده شده است. بدین منظور، از الگوهای مختلف دبی روزهای پیشین (گام تأخیر) برای مدل‌سازی استفاده شد. در مدل‌های استفاده شده در این تحقیق، بهترین الگوی ورودی، الگوی شماره ۵ بود که در آن متغیرهای آبدهی پیشین با پنج گام زمانی تأخیر استفاده شده‌اند و مدل GEP بهترین عملکرد را در پیش‌بینی آبدهی روزانه ایستگاه ملاثانی داشته است.

با توجه به اهمیت مقادیر حداکثر جریان در مدیریت منابع آب، در بخش دیگری از این تحقیق، به بررسی عملکرد مدل‌های مورد بررسی در پیش‌بینی مقادیر حداکثر جریان پرداخته شد. در این خصوص نیز مدل GEP با داشتن متوسط قدر مطلق خطای برابر ۳/۷۳ درصد بهترین عملکرد را در پیش‌بینی مقادیر حداکثر جریان داشت و الگوریتم K-NN و

روش‌های رگرسیون خطی و غیرخطی در اولویت‌های بعدی قرار گرفتند.

نتایج به‌دست آمده در این تحقیق با مطالعه زمانی و همکاران (۲۵) در سرشاخه‌های کارون و نیز امامقلی‌زاده و همکاران (۶) در حوضه آبریز کسیلیان مطابقت دارد. لذا استفاده از مدل برنامه‌ریزی بیان ژن، به‌علت دقت قابل قبول و نیز توانایی برقراری رابطه ضمنی مطلوب بین پارامترهای ورودی و خروجی برای پیش‌بینی میانگین و حداکثر دبی روزانه ایستگاه ملاثانی پیشنهاد می‌شود.

منابع مورد استفاده

1. Aalami, M. T., S. Sadeghfam and M. H. Fazelifard. 2013. Time Series Modeling. Tabriz University. (In Farsi).
2. Aleyasin, A. 2000. Applied River Engineering in Dez and Karun Rivers. Ministry of Energy, Tehran. (In Farsi).
3. Ahani, A. and M. Shourian. 2017. Prediction of monthly streamflow using data driven models . *Journal of Iran-Water Resources Research* 13(2): 207-214. (In Farsi).
4. Azmi, M. and Sh. Araghinejad. 2012. Developed K-nearest neighbor method for river flow prediction. *Journal of Water and Wastewater* 2: 108-119. (In Farsi).
5. Dasarathy, B. V. 1991. Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques. IEEE Computer Society Press, Los Alamitos, CA.
6. Emamgolizadeh, S., R. Karimi Damaneh and H. Mehdipanah. 2016. Estimation of Kasilian watershed runoff by gene expression programming. International Conference on Sustainable Agriculture, Environment and Rural Development, Tehran, Iran. (In Farsi).
7. Ferreira, C. 2001. Gene expression programming: A new adaptive algorithm for solving problems. *Complex Systems* 13(2): 87-129.
8. Govindaraju, R. S. 2000. Artificial neural network in hydrology. *Journal of Hydrologic Engineering* 5(2): 115-123.
9. Güçlü, Y. S. 2018. Multiple Şen-innovative trend analyses and partial Mann-Kendall test. *Journal of Hydrology* 566: 685-704.
10. Karamouz, M. and Sh. Araghinejad. 2011. Advanced Hydrology. 2nd ed., Amirkabir University of Technology, Tehran, Iran. (In Farsi).
11. Khedmati, H., M. Manshouri, M. Heydarizadeh and H. Sedghi. 2010. Zonation and estimation of flood discharge in unguaged sites located in south-east basins of Iran using a combination of flood index and multi-variable regression methods (Sistan and Baluchistan, Kerman, Yazd and Hormozgan provinces). *Journal of Water and Soil* 24(3): 593-906. (In Farsi).
12. Kim, U. and J. J. Kaluarachchi. 2008. Application of parameter estimation and regionalization methodologies to ungauged basins of the Upper Blue Nile River Basin, Ethiopia. *Journal of Hydrology* 362: 39-56.
13. Kisi, O. 2007. Streamflow forecasting using different artificial neural network algorithms. *ASCE, Journal of Hydrologic Engineering* 12(5): 532-539.
14. Kottegoda, N., L. Natale and E. Raiteri. 2004. Some considerations of periodicity and persistence in daily rainfalls. *Journal of Hydrology* 296: 23-37.
15. Omidvarinia, M. 2009. Application of genetic algorithm to optimize the dimensionless input parameters of artificial neural network to estimate suspended load in alluvial rivers. Ph.D Dissertation, Shahid Chamran University, Ahvaz, Iran. (In Farsi).
16. Pande, S., M. McKee and L. A. Bastidas. 2009. Complexity-based robust hydrologic prediction. *Water Resources Research* 45: W10406.
17. Ren, J., B. Ren, Q. Zhang and X. Zheng. 2019. A novel hybrid extreme learning machine approach improved by K nearest neighbor method and fireworks algorithm for flood forecasting in medium and small watershed of loess region. *Water* 11(9): 1848.
18. Sette, S. and L. Boullart. 2001. Genetic programming: Principles and applications. *Engineering Applications of Artificial Intelligence* 14: 727-736.
19. Sharifazari, S. and Sh. Araghinejad. 2013. Develop a non-parametric model to simulate monthly hydrological data. *Journal of Water and Irrigation Management* 1(3): 83-95. (In Farsi).
20. Sharma, A. and U. Lall. 1999. A nonparametric approach for daily rainfall simulation. *Mathematics and Computers in Simulation* 48: 361-371.
21. Soltani, A., M. A. Gorbani, A. Fakheri Fard, S. Darbandi and D. Farsadizadeh. 2011. Genetic programming and its application in rainfall-runoff modeling. *Journal of Water and Soil Science* 20(4): 62-71. (In Farsi).

22. Sun, L., I. Nistor, O. Seidou, S. Sambou, C. M. F. Kebe and S. Tamba. 2013. Prediction of daily discharge at Bakel (Senegal) using multiple linear regression, Kalman filter and artificial neural networks. *In:* 3rd Specialty Conference on Disaster Prevention and Mitigation, DIS-15-1: 9-15.
23. Tsakiri, K., A. Marsellos and S. Kapetanakis. 2018. Artificial neural network and multiple linear regression for flood prediction in Mohawk River, New York. *Water* 10(9): 1158.
24. Wu, C. L. and K. W. Chau. 2010. Data-driven models for monthly streamflow time series prediction. *Engineering Applications of Artificial Intelligence* 23: 1350-1367.
25. Zamani, R., F. Ahmadi and F. Radmanesh. 2015. Comparison of the gene expression programming, nonlinear time series and artificial neural network in estimating the river daily flow (Case study: The Karun river). *Journal of Water and Soil* 28(6): 1172-1182.
26. Zorn, C. R. and A. Y. Shamseldin. 2015. Peak flood estimation using gene expression programming. *Journal of Hydrology* 531: 1122-1128.

Journal of Water and Soil Science
Vol. 25, No. 1, Spring 2021, Isfahan University of Technology, Isfahan, Iran.

Comparison of Gene Expression Programming (GEP) and Parametric and Non-parametric Regression Methods in the Prediction of the Mean Daily Discharge of Karun River (A case Study: Mollasani Hydrometric Station)

M. Alinezhadi, S. F. Mousavi* and Kh. Hosseini[1]

(Received: March 7-2020 ; Accepted: August 5-2020)

Abstract

Nowadays, the prediction of river discharge is one of the important issues in hydrology and water resources; the results of daily river discharge pattern could be used in the management of water resources and hydraulic structures and flood prediction. In this research, Gene Expression Programming (GEP), parametric Linear Regression (LR), parametric Nonlinear Regression (NLR) and non-parametric K- Nearest Neighbor (K-NN) were used to predict the average daily discharge of Karun River in Mollasani hydrometric station for the statistical period of 1967-2017. Different combinations of the recorded data were used as the input pattern to predict the mean daily river discharge. The obtained esults indicated that GEP, with R^2= 0.827, RMSE= 59.45 and MAE= 26.64, had a better performance, as compared to LR, NLR and K-NN methods, at the validation stage for daily Karun River discharge prediction with 5-day lag, at the Mollasani station. Also, the performance of the models in the maximum discharge prediction showed that all models underestimated the flow discharge in most cases.

Keywords: River discharge, Gene expression programming, Linear and nonlinear regression, K- nearest neighbor, Karun.

1- Department of Water Engineering and Hydraulic Structures, Faculty of Civil Engineering, Semnan University, Semnan, Iran.
Corresponding author, Email: fmousavi@semnan.ac.ir