نشریه آب و خاک (علوم و صنایع کشاورزی)
جلد ۳۵، شماره ۲، خرداد – تیر ۱۴۰۰، ص. ۱۶۵-۱۵۳

Journal of Water and Soil
Vol. 35, No. 2, May-June 2021, p. 153-165

مقاله پژوهشی

بررسی کارایی روش‌های پیش‌پردازش داده‌ها در بهبود عملکرد مدل برنامه‌ریزی بیان ژن (مطالعه موردی: رودخانه آب زال)

فرشاد احمدی[1*]

تاریخ دریافت: ۱۳۹۹/۰۲/۰۸
تاریخ پذیرش: ۱۳۹۹/۱۱/۰۹

چکیده

در این مطالعه سعی گردیده تاثیر کاربرد ضرایب فصلی و روش ریاضی تحلیل و پردازش سیگنال تحت عنوان تبدیل موجک در بهبود عملکرد مدل برنامه‌ریزی بیان ژن (GEP) در پیش‌بینی جریان ماهانه رودخانه آب زال در دوره آماری ۱۳۵۱ تا ۱۳۹۶ مورد بحث و بررسی قرار گیرد. بدین منظور داده‌ها در سه حالت مختلف شامل الف) استفاده از داده‌های جریان و در نظر گرفتن نقش حافظه تا چهار تاخیر، ب) دخالت دادن ترم پریودیک در دو حالت خطی (α-GEP) و غیرخطی (PT-GEP) و ج) تجزیه داده‌ها با استفاده از پنج تابع موجک مختلف به دو زیرسری جزئیات و تقریب، آماده و به مدل GEP معرفی گردید. نتایج حاصل از اجرای مدل‌های خطی و غیرخطی GEP نشان داد که در هر دو حالت، مدل با چهار تاخیر به بیشترین دقت در پیش‌بینی جریان رودخانه دست یافته اما عملکرد مدل غیرخطی GEP با توجه به شاخص‌های ارزیابی مورد استفاده اندکی بهتر بود. در مرحله بعد ترم پریودیک به ورودی‌های مدل افزوده شد. براساس نتایج به دست آمده مدل PT-GEP با الگوی M4 کمترین خطا و بیشترین دقت را به خود اختصاص داده و توانسته شاخص RMSE را هشت درصد کاهش دهد. سپس در گام سوم داده‌های جریان رودخانه با استفاده از توابع موجک تجزیه و مدل‌های W-GEP ایجاد گردید. نتایج کلی این پژوهش نشان داد که مدل‌های W-GEP از عملکرد بسیار مطلوبی برخوردار بوده به طوری که می‌توان از آنها به عنوان یک روش موثر در پیش‌بینی جریان میان‌مدت رودخانه‌ها استفاده نمود.

واژه‌های کلیدی: برنامه‌ریزی بیان ژن، تابع موجک، سطح تجزیه، مدل هیبرید

مقدمه

تاکنون به منظور پیش‌بینی جریان رودخانه‌ها روش‌های متعددی نظیر مدل‌های سری زمانی، انواع روش‌های رگرسیونی، مدل‌های توزیعی و نیمه توزیعی مبتنی بر فیزیک حوضه و مدل‌های هوش مصنوعی توسعه یافته‌اند. کاربرد مدل‌های سری زمانی و رگرسیونی به دلیل الزام در رعایت نمودن فرض‌های اولیه آماری نظیر نرمال بودن داده‌ها همواره با چالش‌هایی روبه‌رو بوده است. از طرفی مدل‌های توزیعی و نیمه توزیعی نیازمند اطلاعات فراوانی بوده که بعضاً تهیه و آماده‌سازی آنها با صرف زمان و هزینه‌های زیادی همراه است و در صورت عدم دسترسی به اطلاعات بایستی از مقادیر پیش‌فرض مدل استفاده نمود. اما در سال‌های اخیر استفاده از روش‌های هوش مصنوعی با بهبود عملکرد رایانه‌ها رو به گسترش بوده و از آنها برای حل مسائل پیچیده نظیر پیش‌بینی جریان رودخانه استفاده می‌گردد. یکی از این روش‌ها که در سال‌های اخیر مورد توجه قرار گرفته است روش برنامه‌ریزی بیان ژن[2] می‌باشد. برنامه‌ریزی بیان ژن از خانواده الگوریتم‌های فراتکاملی بوده و توانایی بالایی در مدل‌سازی فرایند-های غیرخطی و پویا داشته و به همین دلیل در مطالعات متعددی مورد استفاده قرار گرفته است که در ادامه به تعدادی از آنها اشاره می‌شود.

مهر و مجدزاده طباطبایی (۱۸) دقت شبکه عصبی مصنوعی و برنامه‌ریزی ژنتیک را در پیش‌بینی جریان رودخانه آبسرده استان لرستان مورد مقایسه قرار دادند. همچنین عملکرد حافظه در دو مدل فوق برای پیش‌بینی دبی جریان بررسی شده و در نهایت نتایج حاصل

۱- 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)
DOI: 10.22067/jsw.2021.14975.0

2- Gene Expression Programming

از دو مدل نشان داد که الگوی ورودی با سه تاخیر بهترین حافظه را داشته و برنامه‌ریزی ژنتیک با دقت بهتری نسبت به شبکه عصبی جریان رودخانه را برآورد می‌نماید. احمدی و همکاران (۲) از دو روش برنامه‌ریزی ژنتیک و ماشین بردار پشتیبان برای پیش‌بینی جریان روزانه رودخانه باراندوزچای در محل ایستگاه هیدرومتری دیزج در خلال سال‌های ۱۳۸۵ تا ۱۳۸۹ بهره بردند. نتایج نشان داد که در هر دو روش، مدل‌های شامل جریان یک، دو و سه روز قبل بالاترین دقت را در مرحله صحت‌سنجی داشتند. با این وجود مقایسه عملکرد دو مدل نشان داد که دقت روش برنامه‌ریزی ژنتیک نسبت به روش ماشین بردار پشتیبان اندکی بیشتر بود. نوحه‌گر و همکاران (۲۱) به منظور پیش‌بینی جریان روزانه رودخانه حوضه آبریز معرف امامه از مدل‌های برنامه‌ریزی ژنتیک و شبکه عصبی مصنوعی استفاده کردند. این محققان گزارش نمودند که روش GP، از میان مدل‌های مختلف به کار رفته، خطای کمتری داشته و به عنوان روش مناسب جهت پیش‌بینی جریان روزانه معرفی گردید. مهر (۱۷) با استفاده از مدل تلفیقی الگوریتم ژنتیک- برنامه‌ریزی بیان ژن جریان ماهانه رودخانه شاویر واقع در حوضه آبریز سفید رود را پیش‌بینی نمودند. نتایج نشان داد که هیبرید GEP و الگوریتم ژنتیک نسبت به مدل ساده GEP از عملکرد بسیار بهتری برخوردار بود. مهر و نورانی (۱۹) با استفاده از روش برنامه‌ریزی بیان ژن چندگانه، فرآیند بارش و رواناب حوضه آبریز هالدیزن[1] واقع در کشور ترکیه را بررسی نموده و گزارش نمودند که مدل مذکور از دقت بسیار بالایی در برآورد جریان برخوردار است. مطالعات دیگری نیز در این زمینه صورت گرفته است که از آن جمله می‌توان به هادی و تامبل (۱۱)، کومار و همکاران (۱۲)، رحمانی رضایی و همکاران (۲۴) و آشفته و همکاران (۴) اشاره نمود.

در کاربرد مدل‌های هوشمند انتخاب جمعیت‌های اولیه تصادفی مختلف و تاثیرگذار در پدیده (که در برنامه‌ریزی بیان ژن به عنوان داده‌های آموزشی از آن‌ها یاد می‌شود) به منظور آموزش فرآیندها و سازوکار حاکم، موجب پیچیدگی الگو و افزایش حافظه درگیر شده و در نتیجه عملکرد مدل را کاهش می‌دهد (۱۸). لذا در پیش‌بینی جریان رودخانه‌ها بایستی داده‌های موثر در اختیار مدل قرار گیرند اما، سری های زمانی ثبت شده هیدرولوژیک، با مشکلاتی همچون خاصیت تناوبی و وجود نویز سفید در داده‌ها مواجه هستند (۱). عوامل مذکور موجب کاهش دقت شده و برآوردها را به حالت ناریب تبدیل می‌نمایند (۲۲). برای حل این مشکلات می‌توان از روش‌هایی همچون پیش‌پردازش داده‌ها و کمک به تشخیص ترم پریودیک[2] (برای مدل‌سازی خاصیت تناوبی) و تلفیق مدل‌های هوشمند و توابع موجک استفاده نمود. استفاده از این روش‌ها در پژوهش‌های مختلفی مورد نظر بوده به طوریکه لوهانی و همکاران (۱۴) و منتصری وقویدل (۲۰) هریک با استفاده از رویکردهای متفاوت خاصیت تناوبی اقدام به مدل‌سازی جریان رودخانه در مقیاس ماهانه نموده و تاثیر مثبت آن را در بهبود عملکرد مدل‌های هوشمند تایید نمودند. تابع موجک نیز یک روش ریاضی بوده که داده‌ها را به دو زیرسری تقریب و جزئیات تقسیم نموده و از این طریق اطلاعات مورد نیاز مدل‌ها را با الگوی مناسب‌تری در اختیار آنها قرار می‌دهد. در مطالعات متعددی همچون پارمار و همکاران (۲۲)، سان و همکاران (۲۶)، فریر و همکاران (۹) و یاسین و همکاران (۲۹) کارایی تلفیق مدل‌های هوشمند و توابع موجک در پیش‌بینی پارامترهای هیدرولوژیک گزارش گردیده است.

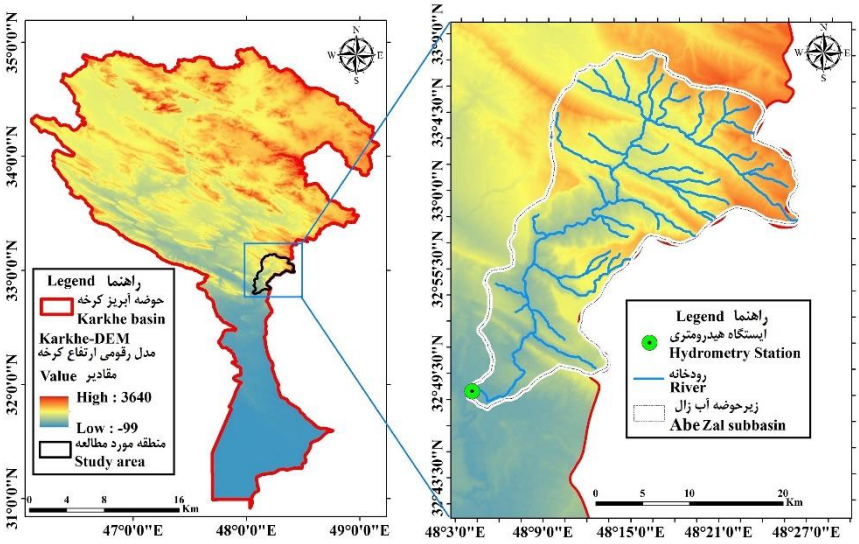
با توجه به پیشینه پژوهش ارائه شده مشاهده می‌شود که تخمین پارامترهای هیدرولوژیکی هم‌چون پیش‌بینی جریان رودخانه‌ها از دیرباز مورد توجه محققین امر بوده و بدین منظور روش‌های متعددی از جمله مدل‌های تجربی- نیمه تجربی، سری‌های زمانی، مدل‌های هوشمند و مدل‌های هیبریدی ارائه شده‌اند. در این میان مدل‌های هیبریدی با تلفیق روش‌های هوشمند و پیش‌پردازش داده‌ها می‌توانند پارامترهای هیدرولوژیک را با دقت قابل توجهی برآورد نمایند. لذا در این تحقیق سعی گردیده است تا میزان بهبود عملکرد روش GEP با استفاده از توابع موجک و افزودن ترم پریودیک مورد بررسی و ارزیابی قرار گرفته و مناسب‌ترین ترکیب هیبریدی برپایه GEP معرفی گردد.

مواد و روش‌ها

داده‌ها و منطقه مورد مطالعه

در این مطالعه از اطلاعات جریان ایستگاه هیدرومتری پل زال واقع بر رودخانه آب زال در دوره آماری ۱۳۵۱ تا ۱۳۹۶ استفاده شده است. رودخانه زال از کوه‌های میان‌کوه شرقی خرم‌آباد لرستان در منطقه‌ای به نام کوس کاوه (باغ کاوه) سرچشمه گرفته و به رودخانه سیمره واقع در منطقه‌ای به نام پل زال (منطقه گرمسیری) می‌ریزد. در طول مسیر (سردسیر) چندین چشمه نیز به این رودخانه می‌ریزند، از جمله: انارود کوچک و بزرگ، چشمه خارماسینه کو و آب چهل گوار (گذر). در ادامه مسیر، این رودخانه از یک تنگه به نام هلد می‌گذرد که انتهای مسیر سردسیر آن محسوب می‌شود. سپس با عبور از دره‌ای عمیق، وارد تنگه کرکی می‌شود. در ادامه با عبور از کول چپ، درکی، دیکپل، تنگه زینکی، تنگه هرندی و تنگه چین زال به پایان راه خود رسیده و به رودخانه سیمره که یکی از سرچمه‌های مهم رودخانه کرخه می باشد می‌پیوندد. شکل ۱ موقعیت ایستگاه و منطقه مورد مطالعه در حوضه آبریز کرخه را نشان می‌دهد. همچنین در جدول ۱ مشخصات فیزیکی و آبدهی حوضه آبریز آب زال ارائه شده است.

1- Haldizen Catchment
2- Periodic Term (PT)



شکل ۱- موقعیت منطقه مورد مطالعه در حوضه آبریز کرخه

Figure 1- Location of study area in the Karkhe basin

جدول ۱- خصوصیات فیزیوگرافی و هیدرولوژیکی حوضه آبریز مورد مطالعه (۱۳۹۶-۱۳۵۱)

Table 1- The physiographic and hydrological characteristics of the study are (1972-2017)

ردیف Row	پارامتر Parameter	مقدار Value
1	میانگین آبدهی در محل ایستگاه هیدرومتری (MCM) Mean annual discharge volume (MCM)	271.5
3	مساحت کل حوضه (km^2) Area (km^2)	618.5
4	محیط کل حوضه (km) Perimeter (km)	135.6
5	طول شاخه اصلی (km) Length of main river (km)	73.8
6	متوسط ارتفاع حوضه (m) Average basin height (m)	1481.5

مدل برنامه‌ریزی بیان ژن

در سال‌های اخیر، سیستم‌های متفاوتی از الگوریتم‌های ژنتیک گسترش یافته‌اند؛ الگوریتم‌های قدرتمندی که از سیر تکاملی طبیعی الهام گرفته‌اند و می‌توانند در طیف وسیعی از علوم بکار گرفته شوند. این الگوریتم‌ها از لحاظ ساختاری می‌توانند به سه گروه اصلی تقسیم شود (۸): ۱) الگوریتم‌های ژنتیک با افرادی شامل کروموزوم‌های خطی با طول ثابت و بدون بیانی پیچیده. در این سیستم‌ها کروموزوم‌ها طبق برتری خصوصیات فردیشان حفظ می‌شوند و به الگوریتم ژنتیک[1] (GA) معروف هستند. ۲) الگوریتم‌های ژنتیک با افرادی شامل ساختارهای شاخه‌ای[2] از اندازه‌ها و اشکال متفاوت و توانا در پذیرش تعداد بیشتری از عوامل. در این سیستم‌ها نیز ساختارهای شاخه‌ای طبق برتری خصوصیات فردیشان حفظ می‌شوند که برنامه‌ریزی ژنتیک (GP) نامیده می‌شوند. ۳) الگوریتم‌های ژنتیک با افراد کدگذاری شده به شکل کروموزوم‌های خطی با طولی ثابت و قابل بیان به شکل ساختارهای شاخه‌ای با اندازه‌ها و اشکال متفاوت. در این سیستم‌ها کروموزوم‌ها بواسطه برتری عوامل سببی روی فنوتیپ[3] (ساختارهای شاخه‌ای) حفظ می‌شوند و به برنامه‌ریزی بیان ژن (GEP) معروف هستند.

این طبقه‌بندی به طور واضح رابطه بین GEP و GP را نشان می‌دهد که هر دو در سیر تکاملی برنامه‌های کامپیوتر به صورت ساختارهای شاخه‌ای بکار گرفته می‌شوند (۸). برنامه‌ریزی بیان ژن

1- Genetic Algorithm
2- Ramified Structures
3- Phenotype

(GEP) نیز همانند الگوریتم ژنتیک (GA) و برنامه‌ریزی ژنتیک (GP)، یک الگوریتم ژنتیکی است بطوری که از جمعیتی از افراد استفاده می‌کند، آنها را مطابق برازندگی انتخاب می‌کند و تغییرات ژنتیکی را با استفاده از یک یا چند عملگر ژنتیکی اعمال می‌نماید. همانطور که ذکر شد، تفاوت اساسی بین این سه الگوریتم، مربوط به ماهیت افراد آنهاست؛ بطوری که در GA افراد رشته‌های خطی با طول ثابت (کروموزوم‌ها) و در GP نهادهای غیرخطی با اندازه‌ها و اشکال متفاوت (درختان تجزیه) می‌باشند؛ در حالی که در GEP، افراد به صورت رشته‌های خطی با طول ثابت (ژنوم یا کروموزوم‌ها) کدگذاری شده و سپس به شکل نهادهای غیرخطی با اندازه‌ها و اشکال متفاوت (یعنی نمایش دیاگرام ساده یا بیان درختی) توصیف می‌شوند (۱۰). فرآیند گام به گام حل یک مسئله با استفاده از برنامه ریزی بیان ژن متشکل از پنج مرحله به شرح زیر می‌باشد: ۱) انتخاب مجموعه ترمینال: که همان متغییرهای مستقل مسئله و متغییرهای حالت سامانه می‌باشند. ۲) انتخاب مجموعه توابع: که شامل عملگرهای حسابی، توابع آزمون و توابع بولی می‌باشد. ۳) شاخص اندازه‌گیری دقت مدل که برمبنای آن می‌توان مشخص نمود که توانایی مدل در حل یک مسئله خاص تا چه اندازه می‌باشد. ۴) مولفه-های کنترل: مقادیر مولفه‌های عددی و متغییرهای کیفی که برای کنترل اجرای برنامه‌ها استفاده می‌شوند. ۵) شرط توقف اجرای برنامه: که معیاری برای حصول نتیجه و توقف اجرای برنامه می‌باشد. در مطالعه حاضر از برنامه GeneXproTools که توسط فریرا (۸) ارئه شده است، برای توسعه و اجرای دو نوع مدل خطی و غیرخطی GEP استفاده شد. مدل خطی GEP براساس عملگرهای $\{+,-,\times,\div\}$ و مدل غیرخطی با به کارگیری توابع $\left\{+,-,\times,\div,\sin,\cos,X^2,X^3,\exp,\sqrt[2]{}\right\}$ توسعه داده شد (۸).

توابع موجک

مطالعات فوریه در قرن ۱۷ میلادی پایه و اساس اولیه تحلیل سیگنال‌ها را مطرح و زمینه را برای ابداع تئوری موجک فراهم نمود. سال‌های متمادی پژوهش‌گران برای تحلیل داده‌های نامنظم و متناوب از تحلیل فوریه استفاده کرده و در بسیاری از موارد نتایج با خطای قابل توجهی همراه بود. برای رفع نقیصه‌های این روش نظریه موجک را معرفی گردید (۲۵). نظریه موجک با کاربرد تئوری‌های مهندسی و ریاضیاتی به سرعت توسعه یافته و سادگی کاربرد آن در زمینه‌های مختلف امکان استفاده و تحلیل داده‌ها را با دقت قابل قبولی ارائه نمود. موج کوچک یا موجک باید دارای تعداد نوسان‌های محدود، بازگشت سریع به صفر در هر دو جهت مثبت و منفی از دامنه خود و میانگین صفر باشد. این خصوصیات تحت عنوان شرط "پذیرفتگی[1]" شناخته می‌شود. یکی از مشخصات کاربردی تابع موجک، الگوریتم فیلترسازی آن است که با عبور دادن داده‌ها از فیلتر مربوطه، آنها را به دو دسته تقریب و جزئیات تقسیم می‌نماید. تقریب نماینده اجزا با فرکانس پایین یا مقیاس بالا و جزئیات شامل اجزایی با مقیاس کوچک یا فرکانس بالا بوده و فرآیند تجزیه امواج می‌تواند تا چندین مرحله ادامه یابد (۲۷). A0، A1، A2 و ... ضرایب بالاترین سطوح تقریب و Dj ضریب جزئیات تابع F(x) است. در صورت ادامه مراحل تجزیه، این روند می‌تواند با تجزیه مداوم تقریب‌ها تکرار گردد. بدین ترتیب یک موج به زیر مجموعه‌هائی مانند شکل ۱ تقسیم می-شود که آن را درخت تجزیه موجک می‌نامند (۲۳).

تابع موجک را می‌توان به عنوان تابعی تعریف نمود که دو ویژگی مهم نوسانی بودن و کوتاه مدت بودن را دارا می‌باشد. $\psi(x)$ تابع موجک است اگر و فقط اگر تبدیل فوریه آن $\psi(\omega)$ رابطه زیر را ارضا نماید:

$$\varphi(0)=\int_{-\infty}^{+\infty}\varphi(x)dx=0 \qquad (۱)$$

شرط همان شرط پذیرفتگی برای موجک $\psi(x)$ بوده و این ویژگی (تابع با میانگین صفر)، چندان محدود کننده نبوده و می‌توان توابع بسیاری را بر اساس آن تابع موجک نامید. $\psi(x)$ تابع موجک مادر است که توابع مورد استفاده در تحلیل، با دو عمل ریاضی انتقال و مقیاس در طول سیگنال مورد بررسی، تغییر اندازه و تغییر محل داده می‌شوند (رابطه ۲).

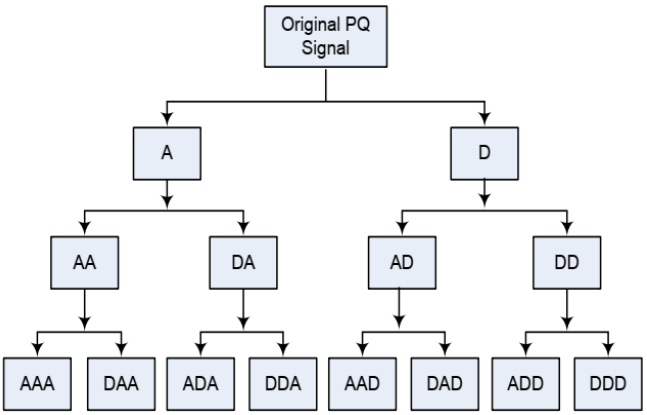
$$\varphi_{a,b}(x)=\frac{1}{\sqrt{a}}\varphi\left(\frac{x-b}{a}\right) \qquad (۲)$$

در نهایت ضرایب موجک در هر نقطه از سیگنال (b) و برای هر مقدار از مقیاس (a) با استفاده از رابطه ۳ قابل محاسبه می‌باشد (۱۶).

$$T(a,b)=\frac{1}{\sqrt{a}}\int_{-\infty}^{+\infty}\varphi\left(\frac{t-b}{a}\right)f(t)dt \qquad (۳)$$

به ازای مقادیر مختلف a و b مقدار T قابل محاسبه بوده و زمانی‌که T بیشترین مقدار مثبت را به خود اختصاص دهد در این صورت می‌توان انتظار داشت که بیشترین انطباق نیز حاصل گردد. در صورتی که $T=0$ باشد، در این صورت انتطباقی وجود نداشته و به ازای Tهای کوچکتر از صفر انطباق عکس یا بیشترین اختلاف وجود خواهد داشت (۶). توابع موجک مادرها دارای انواع بسیاری هستند که در این مطالعه از توابع موجک هار، دابچیز، سیملت، میر و کویفلت برای تجزیه داده‌های ماهانه جریان رودخانه استفاده گردید.

1- Admissibility



شکل ۲- درخت تجزیه موجک

Figure 2- Wavelet packet decomposition tree

یکی از مراحل مهم در کاربرد توابع موجک انتخاب سطح تجزیه مناسب برای تحلیل سیگنال مورد نظر می‌باشد. بدین منظور در مطالعه حاضر از رابطه پیشنهادی دینگ و وانگ (۲۸) به شرح زیر استفاده شد:

$$L = Int[Log(N)] \quad (۴)$$

در رابطه فوق، L تعداد سطح تجزیه، N طول سری داده‌های اولیه و INT عملگر صحیح می‌باشد. با تجزیه هریک از سری داده‌ها، واریانس سری‌های زمانی اولیه به زیرسری‌های تقریب و جزئیات منتقل شده و هیچ گونه اطلاعاتی از سری زمانی اولیه از بین نخواهد رفت (۱۳). بطوری‌که با معکوس نمودن فرایند ارائه شده در شکل ۲ می‌توان با دقت بالایی سری زمانی اولیه را بازسازی نمود (۵). پس از انتخاب سطح تجزیه مناسب می‌توان مدل‌های تلفیقی را ایجاد نمود.

مدل‌های هیبریدی

در این مطالعه برای ایجاد مدل‌های هیبریدی از توابع موجک و خاصیت تناوبی استفاده می‌شود. در اغلب مطالعات انجام شده بیشتر به رابطه خطی موجود در بین داده‌های ورودی به مدل‌ها تاکید شده و رابطه تناوبی چندان مورد توجه قرار نگرفته است از این رو در مطالعه حاضر از عبارت‌های $\cos[2\pi.i/12]$ و $\sin[2\pi.i/12]$ ($i=1,\dots,12$) برای بسط اثر خاصیت پریودیک غیرخطی جریان استفاده شده و به عنوان ورودی به مدل GEP افزوده شده و مدل PT-GEP ایجاد گردید (۱۵). همچنین منتصری و قویدل (۲۰) پیشنهاد نمودند که که یک ضریب تحت عنوان α که نماینده ماه‌های سال می‌باشد و حالت خطی دارد نیز می‌تواند خاصیت تناوبی را برای مدل α-GEP معرفی نماید. در مرحله بعد داده‌های دبی جریان ماهانه با استفاده از توابع موجک مختلف با سطح تجزیه مناسب پردازش شده و به عنوان ورودی به مدل GEP معرفی گردیده و به این صورت مدل W-GEP ایجاد شد.

شاخص‌های ارزیابی مدل

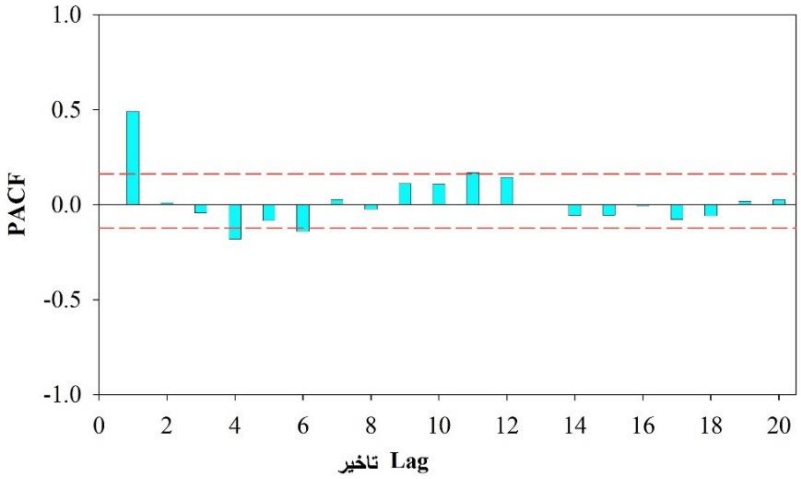
در این مطالعه برای ارزیابی مدل‌های مورد بررسی از معیارهای جذر میانگین مربعات خطا (RMSE)، ضریب همبستگی (R) و میانگین خطای مطلق استفاده شد. لازم به ذکر است مدلی که کمترین مقدار RMSE و MAE و بیشترین مقدار R را به خود اختصاص دهد به عنوان مناسب‌ترین گزینه انتخاب می‌گردد.

$$R = \left(1 - \frac{\sum_{i=1}^{n}(Q_i - \hat{Q}_i)^2}{\sum_{i=1}^{n}(Q_i - \bar{Q})^2}\right)^{0.5} \quad (۵)$$

$$RMSE = \left(\frac{\sum_{i=1}^{n}(Q_i - \hat{Q}_i)^2}{n}\right)^{0.5} \quad (۶)$$

$$MAE = \frac{\sum_{i=1}^{n}(Q_i - \hat{Q}_i)^2}{n} \quad (۷)$$

که در روابط فوق Q_i مقدار مشاهده شده در گام زمانی i ام، $\hat{Q}_i$ مقدار محاسبه شده در همان زمان، n تعداد داده‌ها و $\bar{Q}$ میانگین مقادیر مشاهداتی می‌باشد.



شکل ۳- مقادیر تابع PACF برای سری ماهانه جریان رودخانه

Figure 3- The partial autocorrelation function (PACF) of the monthly streamflow data

جدول ۲- پارامترهای ورودی و خروجی استفاده شده برای توسعه مدل‌های خطی و غیرخطی GEP

Table 2- Input and output parameters used to develop the linear and nonlinear GEP models

ردیف Row	الگو Pattern	اطلاعات ورودی و خروجی Input and output parameters
1	M1	$Q_t = f\ (Q_{t-1})$
2	M2	$Q_t = f\ (Q_{t-1}, Q_{t-2})$
3	M3	$Q_t = f\ (Q_{t-1}, Q_{t-2}, Q_{t-3})$
4	M4	$Q_t = f\ (Q_{t-1}, Q_{t-2}, Q_{t-3}, Q_{t-4})$

نتایج و بحث

پیش‌بینی جریان ماهانه با استفاده از مدل‌های خطی و غیرخطی GEP و ضرایب فصلی

در این مطالعه برای مدل‌سازی جریان ماهانه ۸۰ درصد داده‌ها برای آموزش (از مهر ۱۳۵۱ تا شهریور ۱۳۸۷ به مدت ۴۳۲ ماه) و ۲۰ درصد برای صحت‌سنجی (از مهر ۱۳۸۷ تا شهریور ۱۳۹۶ به مدت ۱۰۸ ماه) در نظر گرفته شد. یکی از مهمترین مراحل مدل‌سازی و پیش‌بینی پارامترهای هیدرولوژیک تعیین ورودی‌های مورد نیاز مدل برای محاسبه لایه خروجی است. به منظور پیش‌بینی جریان رودخانه می‌توان از اطلاعاتی همچون بارش، دما، تبخیر و سایر پارامترهای موثر در پدیده استفاده نمود اما لزوما استفاده از این متغیرها موجب بهبود عملکرد مدل نشده و هزینه‌های جمع‌آوری اطلاعات را نیز افزایش می‌دهد. بنابراین در مطالعات متعددی همچون احمدی و همکاران (۱ و ۳) و تیب و همکارن (۲۷) از داده‌های تاخیر یافته جریان رودخانه برای مدل‌سازی استفاده شده است. اما همواره سوال اصلی در این روش، انتخاب تعداد تاخیرهای مناسب برای مدل‌سازی است. بدین منظور می‌توان از تابع PACF استفاده نمود (۲۰). در شکل ۳ نمودار تابع مذکور برای سری ماهانه ایستگاه پل زال ارائه شده است. با توجه به این شکل مشاهده می‌شود که بعد از تاخیر اول، تاخیرهای دوم و سوم داخل باند اطمینان قرار گرفته و سپس در تاخیر چهارم مقدار PACF خارج از باند اطمینان واقع می‌شود. از این رو در مطالعه حاضر تا چهار تاخیر برای مدل‌سازی در نظر گرفته شده و به شرح جدول ۲ مدل‌های خطی و غیرخطی GEP توسعه داده شدند.

نتایج حاصل از اجرای مدل‌های خطی و غیرخطی GEP در جدول ۳ ارائه شده است. با توجه به این جدول مشاهده می‌شود که در مرحله آموزش مقدار خطای مدل (شاخص‌های RMSE و MAE) در شرایطی که مدل از توابع مختلف استفاده می‌کند کمتر از حالت خطی است. به عنوان مثال مقدار MAE مدل خطی GEP در مرحله آموزش بین ۵/۳۵۸ (m^3/s) تا ۵/۲۵۹ (m^3/s) واقع شده و در مدل غیرخطی این مقدار از ۵/۲۲۰ (m^3/s) تا ۴/۹۳۳ (m^3/s) تغییر می‌کند. به عبارت دیگر کاربرد توابعی مازاد بر توابع بولی این امکان را برای مدل GEP فراهم نموده تا داده‌های مشاهداتی را با خطای کمتری برآورد نماید به طوری که در بهترین حالت خود در الگوی M4 توانسته در مرحله آموزش و تست براساس آماره MAE به ترتیب ۸/۶ ٪ و ۱۶/۶ ٪ خطا را کاهش دهد. بنابراین با توجه به نتایج حاصله، مدل غیرخطی GEP برای ادامه محاسبات انتخاب گردید.

در جدول ۵ نتایج حاصل از اجرای مدل غیرخطی GEP برای ورودی‌های فصلی ارائه شده است. با مقایسه دو جداول ۳ و ۵

مشاهده می‌شود که خطا در حالت فصلی کاهش داشته و در نتیجه استفاده از ضرایب فصلی توانسته دقت مدل GEP را افزایش دهد. همچنین نتایج نشان می‌دهد که با افزایش تاخیر، عملکرد مدل در دو حالت فصلی و غیرفصلی افزایش می‌یابد اما با افزودن تاخیرهای دوم و سوم، بهبود چندانی در مدل‌سازی مشاهده نمی‌شود. دلیل این امر را می‌توان با استفاده از شکل ۳ توجیه نمود. در این شکل PACF مرتبه های دوم و سوم بسیار ناچیز بوده و عملا در باند اطمینان قرار دارند اما تاخیر مرتبه چهار معنی‌دار بوده و توانسته دقت را به طور قابل ملاحظه‌ای در مقایسه با الگوهای M2 و M3 بهبود بخشد.

همانگونه که ذکر گردید، در این مطالعه از ضرایب فصلی برای مدل‌سازی و پیش‌بینی جریان ماهانه رودخانه استفاده گردید. بدین منظور مدل غیرخطی GEP برای سه حالت PT-GEP، α-GEP و PT-αGEP توسعه داده شد. الگوهای ایجاد شده در جدول ۴ ارائه شده است. منتصری و قویدل (۲۰) و لوهانی و همکاران (۱۴) از این شیوه برای ارائه خاصیت تناوبی به مدل‌های هوشمند و از جمله روش GEP استفاده نموده‌اند (۲۰ و ۱۴). این ضرایب می‌توانند با کمی کردن خاصیت تناوبی موجود در داده‌ها که بیشتر مرتبط به زمان می باشد، مرحله آموزش مدل را بهبود بخشند.

نتایج ارائه شده در جدول ۵ نشان می‌دهد که مدل PT-GEP با الگوی M4 توانسته کمترین خطا و بیشترین دقت را در برآورد جریان ماهانه به خود اختصاص دهد و مدل‌های α-GEP و PT-α-GEP در رتبه‌های بعدی قرار می‌گیرند. بنابراین می‌توان چنین استنباط کرد که استفاده از خاصیت تناوبی غیرخطی می‌تواند اطلاعات مفیدتری را در اختیار مدل قرار دهد. اما استفاده همزمان از خاصیت تناوبی خطی (α) و غیرخطی نتوانسته خطای برآورد GEP را کاهش دهد. مهر و همکاران (۱۸) نشان دادند که با افزایش ورودی‌های مدل GP لزوماً نتایج بهتری به دست نمی‌آید. ایشان جریان ماهانه را تا پنج تاخیر زمانی پیش‌بینی نموده و بهترین الگوی ورودی را با سه تاخیر زمانی برای برآورد دبی جریان معرفی نمودند. بنابراین مشاهده می‌شود که اعمال ورودی‌های هم‌نوع همچون دبی یا خاصیت تناوبی نمی‌تواند نتیجه بهتر را به همراه داشته باشد. نتایج این مطالعه با یافته‌های احمدی و همکاران (۱) و فربودنام و همکاران (۷) نیز مطابقت دارد (۱ و ۷).

جدول ۳- مقادیر آماره‌های RMSE، R و MAE حاصل از مدل‌های خطی و غیرخطی GEP

Table 3- Values of RMSE, R and MAE statistics obtained from the linear and nonlinear GEP models

ردیف Row	الگو Pattern	نوع مدل Model type	مرحله آموزش Training phase			مرحله تست Testing phase		
			RMSE (m^3/s)	R	MAE (m^3/s)	RMSE (m^3/s)	R	MAE (m^3/s)
1	M1	Linear	6.405	0.512	5.358	4.657	0.620	3.242
		Non-Linear	6.241	0.601	5.221	4.494	0.624	3.201
2	M2	Linear	6.380	0.603	5.254	4.553	0.622	3.164
		Non-Linear	6.221	0.606	5.143	4.483	0.626	3.178
3	M3	Linear	6.353	0.605	5.159	4.483	0.625	3.147
		Non-Linear	6.166	0.610	5.064	4.451	0.628	2.904
4	M4	Linear	5.970	0.614	5.051	4.307	0.634	2.839
		Non-Linear	**5.531**	**0.623**	**4.933**	**4.093**	**0.660**	**2.782**

جدول ٤- پارامترهای ورودی و خروجی استفاده شده برای توسعه مدل‌های فصلی و غیرخطی GEP

Table 4- Input and output parameters used to develop the seasonal and nonlinear GEP models

ردیف Row	الگو Pattern	نوع مدل Model type	اطلاعات ورودی و خروجی Input and output parameters
1	M1	α-GEP	$Q_t = f\left(Q_{t-1}, \alpha\right)$
		PT-GEP	$Q_t = f\left(Q_{t-1}, \cos[2\pi.i/12], \sin[2\pi.i/12]\right)$
		PT-α-GEP	$Q_t = f\left(Q_{t-1}, \cos[2\pi.i/12], \sin[2\pi.i/12], \alpha\right)$
2	M2	α-GEP	$Q_t = f\left(Q_{t-1}, Q_{t-2}, \alpha\right)$
		PT-GEP	$Q_t = f\left(Q_{t-1}, Q_{t-2}, \cos[2\pi.i/12], \sin[2\pi.i/12]\right)$
		PT-α-GEP	$Q_t = f\left(Q_{t-1}, Q_{t-2}, \cos[2\pi.i/12], \sin[2\pi.i/12], \alpha\right)$
3	M3	α-GEP	$Q_t = f\left(Q_{t-1}, Q_{t-2}, Q_{t-3}, \alpha\right)$
		PT-GEP	$Q_t = f\left(Q_{t-1}, Q_{t-2}, Q_{t-3}, \cos[2\pi.i/12], \sin[2\pi.i/12]\right)$
		PT-α-GEP	$Q_t = f\left(Q_{t-1}, Q_{t-2}, Q_{t-3}, \cos[2\pi.i/12], \sin[2\pi.i/12], \alpha\right)$
4	M4	α-GEP	$Q_t = f\left(Q_{t-1}, Q_{t-2}, Q_{t-3}, Q_{t-4}, \alpha\right)$
		PT-GEP	$Q_t = f\left(Q_{t-1}, Q_{t-2}, Q_{t-3}, Q_{t-4} \cos[2\pi.i/12], \sin[2\pi.i/12]\right)$
		PT-α-GEP	$Q_t = f\left(Q_{t-1}, Q_{t-2}, Q_{t-3}, Q_{t-4} \cos[2\pi.i/12], \sin[2\pi.i/12], \alpha\right)$

جدول ۵- مقادیر آماره‌های RMSE، R و MAE حاصل از مدل‌های فصلی GEP

Table 5- Values of RMSE, R and MAE statistics obtained from the linear and nonlinear GEP models

ردیف Row	الگو Pattern	نوع مدل Model type	مرحله آموزش Training phase			مرحله تست Testing phase		
			RMSE (m^3/s)	R	MAE (m^3/s)	RMSE (m^3/s)	R	MAE (m^3/s)
1	M1	α-GEP	5.894	0.606	5.174	4.238	0.651	3.182
		PT-GEP	5.617	0.616	5.012	4.102	0.702	3.107
		PT-α-GEP	5.738	0.610	5.124	4.213	0.691	3.167
2	M2	α-GEP	5.736	0.612	5.153	4.229	0.659	3.106
		PT-GEP	5.615	0.621	4.983	4.023	0.697	2.921
		PT-α-GEP	5.668	0.618	5.018	4.109	0.684	3.009
3	M3	α-GEP	5.701	0.616	5.127	4.191	0.709	2.957
		PT-GEP	5.523	0.629	4.851	3.956	0.726	2.809
		PT-α-GEP	5.614	0.625	4.964	4.014	0.714	2.907
4	M4	α-GEP	5.510	0.623	5.102	3.882	0.741	2.713
		PT-GEP	**5.411**	**0.642**	**4.691**	**3.789**	**0.808**	**2.697**
		PT-α-GEP	5.456	0.639	4.741	3.940	0.731	2.729

پیش‌بینی جریان ماهانه با استفاده از مدل هیبریدی موجک و GEP (W-GEP)

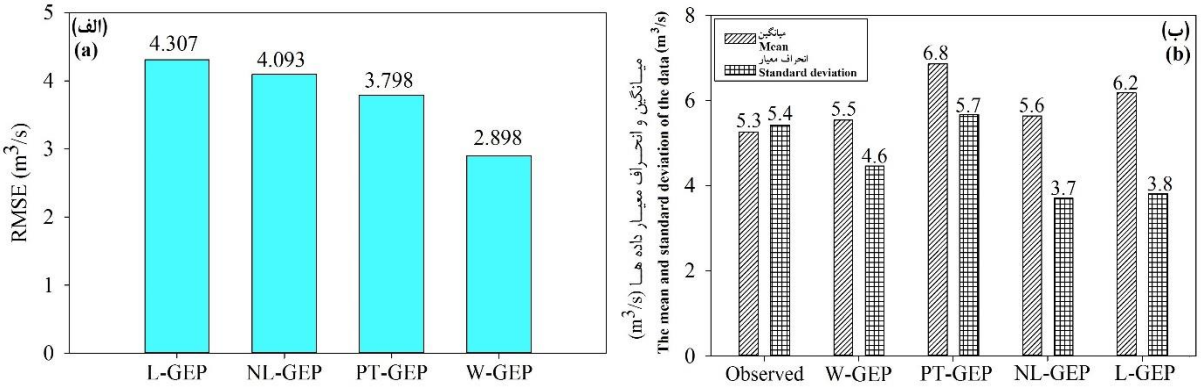
یکی از اهداف مهم مطالعه حاضر کاربرد توابع موجک در پیش پردازش داده‌های ورودی به مدل GEP می‌باشد. در این مرحله داده-های دبی جریان ماهانه توسط موجک‌های هار (haar)، دابچیز ۴ (db4)، سیملت (sym)، میر (mey) و کویفلت (coif) به زیرسری‌های تقریب و جزئیات تبدیل شدند. به این ترتیب تمامی ویژگی‌ها و جزئیات موجود در داده‌ها نمایان شده که این خود باعث افزایش کارایی مدل می‌گردد (۲۵). برای تعیین سطح تجزیه از رابطه ۴ استفاده شد. براین اساس، سطح تجزیه مناسب با توجه به تعداد داده های مورد بررسی عدد دو به دست آمد اما در پژوهش حاضر به منظور اطمینان بیشتر، داده‌ها تا سه سطح تجزیه شدند.

نتایج حاصل از اجرای مدهای W-GEP در جدول ۶ ارائه شده است. با توجه به این جدول مشاهده می‌شود که موجک‌های هار، میر و کویفلت با سطح تجزیه دو توانسته‌اند بهترین اطلاعات را در اختیار مدل GEP قرار دهند. همچنین موجک‌های دابچیز ۴ و سیملت با سطح تجزیه یک به بیشترین دقت دست یافته‌اند. علاوه براین مشاهده می‌شود که هیچ یک از موجک‌ها با سطح تجزیه ۳ نتوانسته کمترین مقدار خطا را حاصل نماید. این امر نشان می‌دهد که رابطه ۴ به خوبی می‌تواند تعداد سطح تجزیه مناسب را تعیین کند. این نتیجه با یافته‌های سلگی و همکاران (۲۵) مطابقت دارد.

جدول ٦- مقادیر آماره‌های RMSE، R و MAE حاصل از مدل هیبریدی W-GEP

Table 6- Values of RMSE, R and MAE statistics obtained from the hybrid W-GEP models

ردیف Row	موجک Wavelet	سطح تجزیه Decomposition level	مرحله آموزش Training phase			مرحله تست Testing phase		
			RMSE (m^3/s)	R	MAE (m^3/s)	RMSE (m^3/s)	R	MAE (m^3/s)
1	haar	1	5.393	0.642	4.817	3.889	0.723	2.825
		2	**4.692**	**0.717**	**4.430**	**3.817**	**0.718**	**2.334**
		3	4.814	0.703	4.664	4.175	0.657	2.696
2	db4	**1**	**4.230**	**0.763**	**4.358**	**3.077**	**0.830**	**1.976**
		2	4.400	0.743	4.669	3.100	0.826	2.155
		3	4.168	0.764	4.627	3.810	0.748	2.580
3	sym	**1**	**4.003**	**0.770**	**4.116**	**2.898**	**0.847**	**1.745**
		2	4.326	0757	4.593	2.976	0.840	2.053
		3	4.286	0.761	4.665	3.103	0.825	2.222
4	mey	1	4.826	0.702	4.895	3.762	0.746	2.720
		2	**4.291**	**0.746**	**4.494**	**3.192**	**0.811**	**2.151**
		3	4.373	0.742	4.589	3.210	0.808	2.183
5	coif	1	4.980	0.703	4.717	3.881	0.721	2.839
		2	**4.785**	**0.705**	**4.693**	**3.569**	**0.760**	**2.559**
		3	4.796	0.701	4.723	3.584	0.751	2.631



شکل ٤- (الف) مقادیر آماره RMSE و (ب) میانگین و انحراف معیار داده‌های مشاهداتی و محاسباتی حاصل از مدل‌های مورد بررسی

Figure 4- (a) The values of RMSE statistics, (b) the mean and standard deviation of the observed and estimated data obtained from considering models

بهترین عملکرد مدل W-GEP با موجک سیملت و سطح تجزیه یک بوده است که بیشترین دقت و کمترین خطا را در بین کلیه مدل های اجرا شده اعم از خطی و غیرخطی، فصلی و غیرفصلی و هیبرید موجکی داشته است.

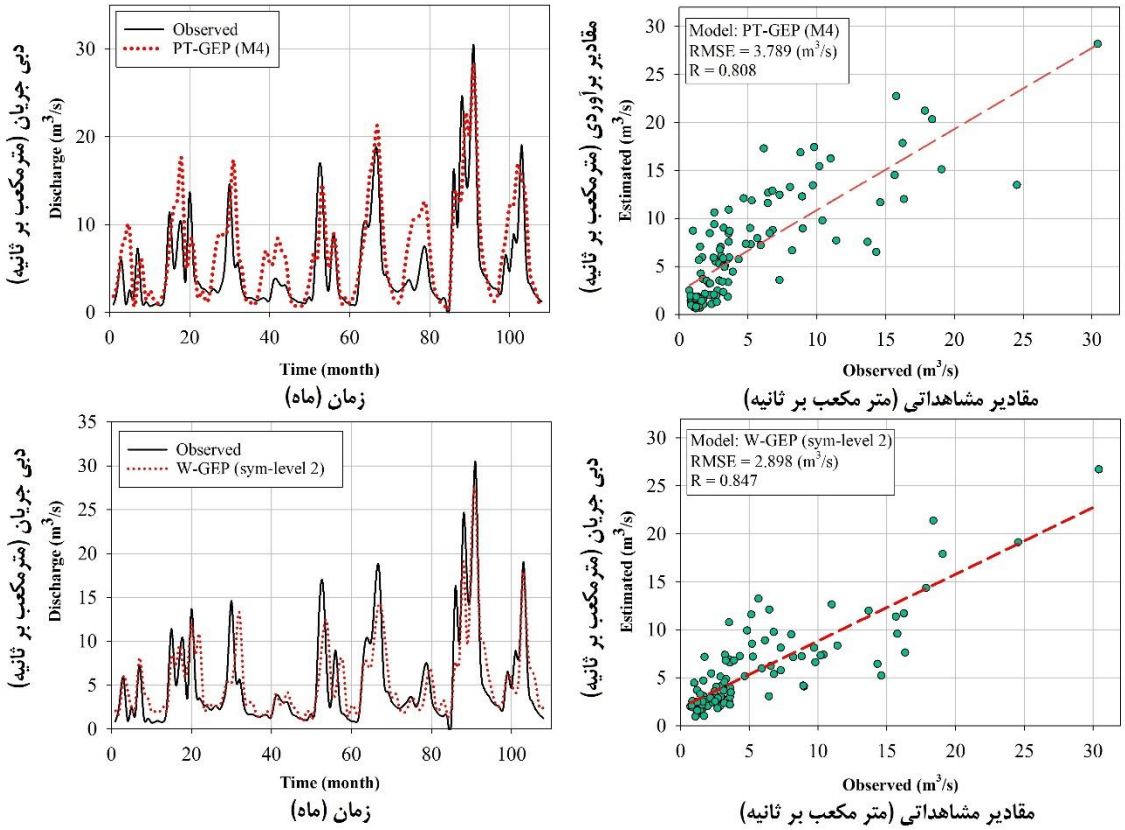
در شکل ۴- الف نمودار ستونی مقادیر آماره RMSE برای کلیه مدل‌های مورد استفاده در بهترین حالت خود در مرحله تست ارائه شده است. در این شکل به خوبی می‌توان تاثیر پیش‌پردازش داده‌ها را در بهبود عملکرد مدل GEP مشاهده نمود. به طوریکه مدل W-GEP نسبت به مدل‌های L-GEP، NL-GEP و PT-GEP توانسته است مقدار آماره RMSE را به ترتیب ۴۸/۶ درصد، ۴۱/۲ درصد و ۳۱/۱ درصد کاهش دهد. این امر به خوبی نشان می‌دهد که روش توابع موجک از توانایی بسیار بالایی در پیش‌پردازش داده‌ها برخوردار بوده و دقت عملکرد مدل‌های هوشمند را می‌توانند به طرز چشم‌گیری افزایش دهند. یکی دیگر از روش‌های مهم ارزیابی عملکرد مدل‌ها در پیش‌بینی پارامترهای هیدرولوژیک، مقایسه میانگین و انحراف معیار داده‌های مشاهداتی و محاسباتی می‌باشد. در شکل ۴- ب میانگین و انحراف معیار داده‌های مشاهداتی و محاسباتی حاصل از مدل‌های مختلف در مرحله تست نشان داده شده است. با توجه به این شکل مشاهده می‌شود که مدل W-GEP با اختلاف کمی میانگین و انحراف داده‌ها را (به ترتیب با تفاوت ۰/۲ و ۰/۸ متر مکعب در ثانیه) در مرحله تست برآورد نموده است. در این بین مدل L-GEP بیشترین اختلاف را با داده‌های مشاهداتی داشته است. برای مدل PT-GEP نیز مقدار میانگین داده‌های محاسباتی در حدود ۱/۵ متر مکعب در ثانیه بیشتر از مقادیر مشاهداتی بوده اما انحراف معیار آن بسیار نزدیک است که می‌تواند به دلیل عدم توانایی مدل در برآورد مقادیر جریان‌های کمینه باشد. برای بررسی بیشتر این موضوع می توان از نمودارهای گرافیکی و پراکندگی جریان مشاهداتی و تخمینی استفاده نمود.

شکل ۵، نمودارهای گرافیکی و پراکندگی مدل‌های W-GEP و PT-GEP را در مرحله آزمون برای بهترین حالت نشان می‌دهد. با توجه به این شکل مشاهده می‌شود که مدل PT-GEP مقادیر جریان کمینه را باخطای بیشتری نسبت به W-GEP پیش‌بینی نموده و مقادیر دبی پیک نیز در هر دو مدل با خطا برآورد شده است.

نتیجه‌گیری

در پژوهش حاضر سعی گردید تا با استفاده از پیش‌پردازش داده‌ها با روش‌های ضرایب فصلی خطی، ضرایب فصلی غیرخطی و توابع موجک عملکرد مدل برنامه‌ریزی بیان ژن (GEP) در پیش‌بینی جریان ماهانه رودخانه آب زال بهبود یابد. نتایج حاصل از پژوهش حاضر به شرح زیر قابل ارائه می‌باشد:

- نتایج به دست آمده نشان می‌دهد که روش خطی GEP (L-GEP) که در آن برای فرآیند مدل‌سازی فقط از توابع بولی استفاده می‌گردد، در مقایسه با مدل غیرخطی GEP (NL-GEP)، از عملکرد ضعیف‌تری در برآورد جریان رودخانه برخوردار می‌باشد. هر دو مدل L-GEP و NL-GEP با توالی چهار ماهه دبی جریان (الگوی M4) به بیشترین دقت دست یافته اما عملکرد چندان مطلوبی را در پیش‌بینی جریان ماهانه از خود نشان ندادند.



شکل ۵- مقادیر مشاهداتی و محاسباتی حاصل از مدل‌های PT-GEP و W-GEP در مرحله آزمون

Figure 5- The values of the observed and estimated data obtained from PT-GEP and WGEP models in test phase

❖ نتایج حاصل از افزودن ضرایب فصلی به عنوان ورودی به مدل NL-GEP 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 جمع آوری 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. استفاده از این اطلاعات ممکن است موجب درگیر کردن حافظه مدل شده و در نهایت عملکرد مدل را دچار اختلال نماید بنابراین پیشنهاد می‌گردد که در مطالعات آتی نیز تاثیر کاربرد ترم پریودیک با مدل‌ها و پارامترهای مختلف در بهبود نتایج مدل‌سازی مورد ارزیابی قرار گیرد.

❖ در این مطالعه علاوه بر روش‌های ساده ذکر شده از یک روش ریاضی پیچیده مبتنی بر تجزیه سیگنال تحت عنوان توابع موجک برای بهبود عملکرد مدل GEP استفاده شد. نتایج نشان داد که استفاده از این روش خطای مدل را تا ۴۸/۶ درصد کاهش داده و به خوبی می‌تواند میانگین و واریانس داده‌ها را مدل‌سازی نماید. بهترین عملکرد مدل W-GEP با تابع موجک سیملت در سطح دوم تجزیه به دست آمده و مقادیر آمارهای RMSE، MAE و R در مرحله آزمون به ترتیب برابر با ۲/۸۹۸ متر مکعب در ثانیه، ۱/۷۴۵ متر مکعب در ثانیه و ۰/۸۴۷ محاسبه گردید.

سپاسگزاری

بدین‌وسیله از حمایت مالی معاونت پژوهش و فناوری دانشگاه شهید چمران اهواز در قالب پژوهانه (GN: SCU.WH98.44291) در انجام این تحقیق تشکر و قدردانی می‌گردد.

منابع

1- Ahmadi F., Dinpashoh Y., Fakheri F. A., Khalili K., and Darbandi S. 2015. Comparing nonlinear time series

models and genetic programming for daily river flow forecasting (Case study: Barandouz-Chai River). Journal of Water and Soil Conservation 22(1) : 121-169. (In Persian with English abstract)
2- Ahmadi F., Radmanesh F., and Mirabbasi Najaf Abadi R. 2014. Comparison between Genetic Programming and Support Vector Machine Methods for Daily River Flow Forecasting (Case Study: Barandoozchay River). Journal of Water and Soil 28(6): 1162-1171. (In Persian with English abstract)
3- Ahmadi F., Radmanesh F., and Mirabbasi R. 2016. Comparing the performance of Support Vector Machines and Bayesian networks in predicting daily river flow (Case study: Baranduz Chai River). Journal of Water and Soil Conservation 22(6): 171-186. (In Persian with English abstract)
4- Ashofteh P.S., Bozorg-Haddad O., and Loáiciga H.A. 2020. Logical genetic programming (LGP) application to water resources management. Environmental Monitoring and Assessment 192(1): 34-42.
5- Daubechies I. 1992. Ten lectures on wavelets. 2nd ed. Philadelphia: SIAM, CBMS-NSF regional conference series in applied mathematics 61.
6- Deka P.C., and Prahlada R. 2012. Discrete wavelet neural network approach in significant wave height forecasting for multistep lead time. Ocean Engineering 43: 32-42.
7- Farbodfam N., Ghorbani M.A., and Aalami M.T. 2009. Forecasting river flow using genetic programming (Case study: Lighwan watershed). Journal of Water and Soil Science 19(1):107-123. (In Persian with English abstract)
8- Ferreira C. 2002. Genetic representation and genetic neutrality in gene expression programming. Advances in Complex Systems 5(4): 389-408.
9- Freire P.K.D.M.M., Santos C.A.G., and da Silva G.B.L. 2019. Analysis of the use of discrete wavelet transforms coupled with ANN for short-term streamflow forecasting. Applied Soft Computing 80: 494-505.
10- Ghorbani M.A., Shiri J., and Kazemi H. 2010. Estimation of Maximum, Mean and Minimum Air Temperature in Tabriz City Using Artificial Intelligent Methods. Journal of Agriculture Science 20(4): 87-104. (In Persian with English abstract)
11- Hadi S.J., and Tombul M. 2018. Monthly streamflow forecasting using continuous wavelet and multi-gene genetic programming combination. Journal of Hydrology 561: 674-687.
12- Kumar M., and Sahay R.R. 2018. Wavelet-genetic programming conjunction model for flood forecasting in rivers. Hydrology Research 49(6): 1880-1889.
13- Labat D. 2005. Recent advances in wavelet analyses: Part 1. A review of concepts. Journal of Hydrology 314: 275-288.
14- Lohani A.K., Kumar R., and Singh R.D. 2012. Hydrological time series modeling: A comparison between adaptive neuro-fuzzy, neural network and autoregressive techniques. Journal of Hydrology 442: 23-35.
15- Mallat S.G. 1998. A wavelet tour of signal processing, San Diego.
16- Grossmann A., and Morlet J. 1984. Decomposition of Hardy functions into square integrable wavelets of constant shape. SIAM Journal on Mathematical Analysis 15(4): 723-736.
17- Mehr A.D. 2018. An improved gene expression programming model for streamflow forecasting in intermittent streams. Journal of Hydrology 563: 669-678.
18- Mehr A.D., and Majdzadeh Tabatabai M.R. 2010. Prediction of Daily Discharge Trend of River Flow Based on Genetic Programming. Journal of Water and Soil 24(2): 325-333. (In Persian with English abstract)
19- Mehr A.D., and Nourani V. 2018. Season algorithm-multigene genetic programming: a new approach for rainfall-runoff modelling. Water Resources Management 32(8): 2665-2679.
20- Montaseri M., and Zamanzad Ghavidel S. 2014. River Flow Forecasting by Using Soft computing. Journal of Water and Soil 28(2): 394-405. (In Persian with English abstract)
21- Nohegar A., Motamednia M., and Malekian A. 2016. Daily river flood modeling using genetic programming and artificial neural network (Case study: Amameh representative watershed). Physical Geography Research Quarterly 48(3): 367-383. (In Persian with English abstract)
22- Parmar K. S., Makkhan S.J.S., and Kaushal S. 2019. Neuro-fuzzy-wavelet hybrid approach to estimate the future trends of river water quality. Neural Computing and Applications 31(12): 8463-8473.
23- Polikar R. 1996. Fundamental concepts and an overview of the wavelet theory. Second Edition, Rowan University, College of Engineering Web Servers, Glassboro. NJ. 08028.
24- Rahmani-Rezaeieh A., Mohammadi M., and Mehr A.D. 2020. Ensemble gene expression programming: a new approach for evolution of parsimonious streamflow forecasting model. Theoretical and Applied Climatology 139(2): 549-564.
25- Solgi A., Zarei H., and Golabi M. 2017. Performance Assessment of Gene Expression Programming Model Using Data Preprocessing Methods to Modeling River Flow. Journal of Water and Soil Conservation 24(2): 185-201. (In Persian with English abstract)
26- Sun Y., Niu J., and Sivakumar B. 2019. A comparative study of models for short-term streamflow forecasting with emphasis on wavelet-based approach. Stochastic Environmental Research and Risk Assessment 33(10): 1875-1891.
27- Tayyab M., Zhou J., Dong X., Ahmad I., and Sun N. 2019. Rainfall-runoff modeling at Jinsha River basin by

integrated neural network with discrete wavelet transform. Meteorology and Atmospheric Physics 131(1): 115-125.

28- Wang W., and Ding J. 2003. Wavelet Network Model and Its Application to the Prediction of Hydrology. Nature and Science, pp. 67-71.

29- Yaseen Z.M., Sulaiman S.O., Deo, R.C., and Chau K.W. 2019. An enhanced extreme learning machine model for river flow forecasting: State-of-the-art, practical applications in water resource engineering area and future research direction. Journal of Hydrology 569: 387-408.

Journal of Water and Soil
Vol. 35, No. 2, May-June 2021, p. 153-165

نشریه آب و خاک (علوم و صنایع کشاورزی)
جلد ۳۵، شماره ۲، خرداد – تیر ۱۴۰۰، ص. ۱۶۵-۱۵۳

Evaluation of the Efficiency of Data Preprocessing Methods on Improving the Performance of Gene Expression Programming Model (Case Study: Ab Zal River)

F. Ahamdi[1*]

Received: 27-04-2020
Accepted: 28-01-2021

Introduction: Surface water has always been one of the most essential pillars of water projects and, with modeling and predicting the river flow, in addition to the management and utilization of water resources, it is possible to inhibit the natural disasters such as drought and floods. Therefore, researchers have always tried to improve the accuracy of hydrological parameters estimation by using new tools and combining them. In this study, the effect of seasonal coefficients and mathematical methods of signal analysis and signal processing on wavelet transform to improve the performance of the Gene Expression Programming (GEP) model were discussed.

Materials and Methods: In the present study, for the prediction of the monthly flow of Ab Zal River, the information of Pol Zal hydrometric station in period 1972 to 2017 was used. In the next step, different input patterns need to be ready. To this purpose, the data are presented in three different modes: (a) the use of flow data and considering the role of memory up to four delays; (b) the involvement of the periodic term in both linear (?-GEP) and nonlinear (PT-GEP) states, and (c): data analysis using the Haar wavelet, Daubechies 4 (db4), Symlet (sym), Meyer (mey), and Coiflet (coif), was done in two subscales, prepared, and introduced to the GEP model. To better analyze the effect of mathematical functions used in the GEP method, two linear modes (using Boolean functions including addition, multiplication, division, and minus) and nonlinear (including quadratic functions, etc.) were considered. The wavelet transform is a powerful tool in decomposing and reconstructing the original time series. Wavelet function is a type of function that has an oscillating property and can be quickly attenuated to zero. Modeling was done based on 80% of recorded data (432 months) and the validation was done based on the remaining 20% (108 months). To evaluate the performance of each of models, statistical indices such as mean square error (RMSE), mean absolute error (MAE), and correlation coefficient (R) were used.

Results and Dissection: The results of linear and nonlinear GEP models showed that in both cases, the four-delay model achieved the most accuracy in river flow prediction. Still the performance of nonlinear GEP model according to RMSE (4.093 (m3/s)), MAE (2.782 (m3/s)) and R (0.660) were better than another, respectively. In the next step, the periodic term was added to the model inputs. Based on the results, the PT-GEP model with M4 pattern had the lowest error, the highest accuracy and was able to reduce the RMSE index by 8%. Then, in the third step, the river flow data were divided into approximate subdivisions and details using five wavelet functions. The most appropriate level of analysis based on the number of data was considered as number three. The results of the W-GEP modes showed an excellent performance of this method so that the model was able to reduce the RMSE statistics with 48.6%, 41.2%, and 31.1% compared to the L-GEP, NL-GEP and PT-GEP methods, respectively. Also, the best performance of the W-GEP model with the Symlet wavelet and the decomposition level of one had the highest accuracy (R=0.847) and the lowest error (RMSE =2.898 (m3/s) and MAE =1.745 (m3/s) among all models (35 models) such as linear and nonlinear, seasonal and non-seasonal and wavelet hybrid models.

Conclusion: Based on the results, it can be concluded that the overall use of data preprocessing methods (including seasonal coefficients and wavelet functions) has improved the performance of the GEP model. However, the combination of wavelet functions with the GEP model has significantly increased the accuracy of the modeling. Therefore, it is recommended as the most suitable tool for river flow forecasting.

Keywords: Decomposition level, Gene expression programming, Hybrid model, Wavelet function

1- Assistant Professor, Department of Hydrology and Water Resources, Shahid Chamran University of Ahvaz, Ahvaz, Iran
(*- Corresponding Author Email: F.ahmadi@scu.ac.ir)
DOI: 10.22067/jsw.2021.14975.0