نشریه آبیاری و زهکشی ایران
شماره۳، جلد ۱۵، مرداد - شهریور ۱۴۰۰، ص. ۵۹۷-۵۸۲

Iranian Journal of Irrigation and Drainage
No. 3, Vol. 15, Aug.-Sep. 2021, p. 582-597

مقاله علمی - پژوهشی

بررسی تأثیر پیش‌پردازش داده‌ها و عوامل مدل‌سازی برنامه‌ریزی بیان ژن در دقت پیش‌بینی سری‌های زمانی

مریم صالحی[۱]، سید احسان فاطمی[۲*]

تاریخ دریافت: ۱۳۹۹/۱۰/۳۰ تاریخ پذیرش: ۱۴۰۰/۰۱/۲۴

چکیده

سری زمانی هیدرولوژیک عاملی وابسته به زمان است که یافتن نحوه تغییرات و پیش‌بینی آن مهم‌ترین هدف تجزیه‌وتحلیل سری‌های زمانی است. هدف این تحقیق بررسی هم‌زمان خصوصیات سری زمانی و پیش‌پردازش آن‌ها و پارامترهای مهم مدل برنامه‌ریزی بیان ژن جهت پیش‌بینی‌های با دقت بالا در مراحل آموزش و صحت‌سنجی است. در این پژوهش ازسری‌های زمانی عمق آب زیرزمینی ایستگاه دشت چمچمال واقع در استان کرمانشاه با دوره زمانی ۱۲ساله و اقلیم کوهستانی و سری زمانی ماهانه دمای آلاسکا با دوره زمانی ۵۰ساله و اقلیم سرد و خشک استفاده‌شده است. برای مدل-سازی سری‌های زمانی مذکور از نرم‌افزار Genexprotools5.0 استفاده‌شده است.

نتایج این تحقیق، نشان داد تناوبی بودن خصوصیات داده موجود در سری زمانی دما، سبب بروز نتایج همبستگی بالای ۹۰٪ در مراحل مختلف آموزش و صحت‌سنجی گردید به‌طوری‌که اثر پارامترهای مختلف بیان ژن کمتر از ۱۰درصد در بهبود نتایج است. از سوی دیگر با بررسی سری زمانی عمق آب زیرزمینی که فاقد خصوصیت تناوبی و دارای شکلACF نزولی است، نتایج پیش‌بینی مدل GEP با هر پارامتر تأثیرگذار بیان ژن، R بیش از ۴۴٪ در مرحله صحت‌سنجی حاصل نشد. این بدان معنی است که پیش‌پردازش سری زمانی اثرگذاری بیشتری در نتایج پیش‌بینی دارد. به‌طوری‌که با حذف‌ترم تناوب نتایج پیش‌بینی در همه مراحل مدل‌سازی به طرز معنی‌داری کاهش می‌یابد. در این حالت بهترین R برای قسمت صحت‌سنجی ۵۰ درصد است.

واژه‌های کلیدی: بیان ژن، پیش‌بینی، پیش‌پردازش، تناوب، سری زمانی

مقدمه

داده‌های هواشناسی و هیدرولوژیکی به‌عنوان اطلاعات‌پایه و اساسی در طراحی و مدیریت پروژه‌های منابع آب به کار می‌رود. بررسی روند و ایستایی در سری‌های زمانی هیدرولوژیکی می‌تواند در تفسیر رابطه بین فرآیندهای هیدرولوژیکی و تغییرات محیطی در مناطق موردمطالعه کمک مؤثری داشته باشد (Salas et al., 1996). برنامه‌ریزی ژنتیک یک روش پرکاربرد در منابع آب و هیدرولوژی است. الگوریتم‌های تکاملی که برنامه‌ریزی ژنتیک نیز عضوی از آن‌ها هستند توانایی الگوبندی فرآیندهای کاملاً غیرخطی و پویا رادارند (زمانی و همکاران،۱۳۹۳). روش برنامه‌ریزی بیان ژن، در سال۱۹۹۹ توسط فریرا ابداع شد. این روش ترکیبی از روش‌های GP[3] و GA[4] بوده که در آن، کروموزوم‌های خطی و ساده با طول ثابت، مشابه با آنچه در الگوریتم ژنتیک استفاده می‌شود و ساختارهای شاخه‌ای با اندازه‌ها و اشکال متفاوت، مشابه با درختان تجزیه در برنامه‌ریزی ژنتیک، ترکیب می‌شوند و ازآنجایی‌که در این روش تمام ساختارهای شاخه‌ای با اندازه‌ی اشکال متفاوت، در کروموزوم‌های خطی با طول ثابت کدگذاری می‌شوند، معادل این است که در این روش فنوتیپ و ژنوتیپ از هم جدا می‌شوند و سیستم می‌تواند از تمام مزایای تکاملی به سبب وجود آن‌ها بهره‌مند شود. اکنون باوجوداینکه فنوتیپ

۱- کارشناس ارشد مهندسی منابع آب، پردیس کشاورزی و منابع طبیعی، دانشگاه رازی، کرمانشاه، ایران
۲-استادیار گروه مهندسی آب، پردیس کشاورزی و منابع طبیعی، دانشگاه رازی، کرمانشاه، ایران
(* - نویسنده مسئول: (Email: e_fatemi78@yahoo.com
DOR: 20.1001.1.20087942.1400.15.3.9.3

3- Genetic programming
4- Genetic algorithm

در GEP[1]، همان نوع از ساختارهای شاخه‌ای مورداستفاده درGP را شامل می‌شود، اما ساختارهای شاخه‌ای که به‌وسیله GEP استنتاج می‌شوند (که بیان درختی نیز نامیده می‌شود) مبین تمامی ژنوم‌های مستقل هستند (Lopes and Weinert, 2004). بنا به اهمیت موضوع، تاکنون پژوهشگران مختلفی در سراسر جهان به مطالعه سری‌های زمانی هیدرولوژیکی ازجمله رودخانه‌ها به کمک برنامه‌ریزی بیان ژن پرداخته‌اند (زمانی و همکاران،۱۳۹۳).

روش برنامه‌ریزی بیان ژن یک روش مناسب و علمی در پیش-بینی روابط بارش رواناب است. مطالعه پدیده حمل رسوب در آبراهه‌ها و استفاده از روش برنامه‌ریزی ژنتیک یک رهیافت مناسب برای شبیه‌سازی بار معلق است (Aytek and Kisi., 2008). شری و کیس به‌منظور بررسی نوسانات کوتاه‌مدت سطح آب زیرزمینی دو چاه در ترکیه، از روش‌های برنامه‌ریزی بیان ژن و فازی- عصبی استفاده کردند. نتایج به‌دست‌آمده از پژوهش ایشان بیانگر مناسب بودن دو روش بالا در بررسی نوسانات سطح ایستابی بود (Shiri and Kisi., 2011). تراور و گاون از مدل برنامه‌ریزی بیان ژن برای تخمین تبخیر و تعرق در منطقه‌ای در آفریقا پرداختند و دقت این مدل را قابل‌قبول گزارش نمودند (Traore and Guven, 2012).

آزیمتالا و زهیری برای پیش‌بینی دبی جریان در مقاطع مرکب از دو روش برنامه‌ریزی بیان ژن و مدل درختیM5[2] استفاده کردند. نتایج تحقیقات نشان داد که هرچند هردو مدل از دقت بالایی برخوردار بودند، اما دقت روش برنامه‌ریزی ژنتیک از مدل درختیM5 بالاتر بود (Azamathulla and Zahiri, 2012)

نتایج بررسی عملکرد روش برنامه‌ریزی بیان ژن را درروند یابی سیلاب رودخانه زنگمار در مقایسه با روش موج دینامیکی نشان‌داد برنامه‌ریزی بیان ژن قادر است با دقت بیشتری حجم آب نمود خروجی را پیش‌بینی کند. اما الگوی موج دینامیک در خصوص دبی اوج و زمان وقوع آن برتری دارد(قبادیان و همکاران،۱۳۹۲). نتایج حاصل از بررسی پیش‌بینی عمق سطح آب دریاچه ارومیه حاکی از دقت مطلوب برنامه‌ریزی بیان ژن در شبیه‌سازی نوسانات سطح آب است (کاوه کار و همکاران،۱۳۹۲). نتایج به‌دست‌آمده از مقایسه مدل-های فازی عصبی، شبکه عصبی، منحنی سنجه و بیان ژن جهت تخمین میزان رسوبات معلق رودخانه آجی چای واقع در آذربایجان شرقی نشان‌دهنده عملکرد بهتر مدل برنامه‌ریزی بیان ژن در مقایسه با سایر مدل‌ها است (ثانی خانی و همکاران،۱۳۹۴).

امامقلی زاده و همکاران (۱۳۹۵) بارش رواناب حوضه کسیلیان را به کمک روش برنامه‌ریزی بیان ژن و با استفاده از پارامترهای هواشناسی و هیدرولوژیکی حوضه، مدل‌سازی کردند. نتایج بیانگر توانایی این روش در پیش‌بینی بارش رواناب حوضه مذکور است. پیش‌بینی نقطه شبنم روزانه با استفاده از برنامه‌ریزی بیان ژن (GEP) نیز نشان‌داد که با داده‌های ثبت‌شده در دو ایستگاه رودخانه لیقوان چای در آذربایجان شرقی و ایستگاه آب‌سنجی ونیار در آبخیز آجی چای و روش GEP، به‌خوبی می‌توان نقطه شبنم روزانه را پیش‌بینی کرد (مهدی زاده ،۱۳۹۶).

ادیب و همکاران جریان رودخانه کسیلیان را با استفاده برنامه-ریزی بیان ژن پیش‌بینی کردند. نتایج نشان داد که روش‌های مختلف نرمال‌سازی و حذف روند داده‌ها و حذف خصوصیات تناوبی داده‌ها به‌منظور ایستایی سری زمانی، نتایج نرم‌افزار را بهبود می‌بخشد و درنهایت معیارهای آماری قابل‌قبولی برای پیش‌بینی دبی روزانه حاصل شد (Adib et al., 2017). بررسی عملکرد روش برنامه‌ریزی بیان ژن در پیش‌بینی تابش خورشیدی روزانه در گستره ایران بر اساس داده‌های هواشناسی در ۳۱ ایستگاه در گستره ایران نشان داد روش GEP دقیق‌ترین نتایج را در تخمین تابش خورشیدی روزانه در گستره ایران دارد (خادم پور و همکاران،۱۳۹۷).

سلگی و همکاران (۱۳۹۷) در مدل‌سازی جریان روزانه و ماهانه رودخانه گاماسیاب از مدل برنامه‌ریزی بیان ژن استفاده کردند. برای افزایش عملکرد مدل از دو روش پیش‌پردازش داده‌ها یعنی تبدیل موجک و تجزیه به مؤلفه‌های اصلیPCA استفاده کردند. نتایج نشان داد که عملکرد مدل GEP در دوره ماهانه عملکرد کاهش‌یافته است. آن‌ها با مقایسه مدل ترکیبی برنامه‌ریزی بیان‌ژن-موجک با مدل برنامه‌ریزی بیان ژن نشان داند که عملکرد مدل ترکیبی در هر دو دوره زمانی روزانه و ماهانه از مدل ساده بهتر بوده است. علی دادی ده‌کهنه و همکاران (۱۳۹۸) به ارزیابی دو مدل ژنتیکی برنامه‌ریزی ژنتیک و برنامه‌ریزی بیان ژن با استفاده از داده‌های روزانه جریان، دما، بارش و تبخیر در ایستگاه تله‌زنگ جهت مدل‌سازی جریان رودخانه دز پرداختند. نتایج نشان داد که مدل برنامه‌ریزی بیان ژن دارای عملکرد بهتری است. علاوه بر این، سرعت اجرای مدل برنامه‌ریزی بیان ژن نسبت به مدل برنامه‌ریزی ژنتیک بیشتر بوده و در زمان کمتری قادر به ارائه نتایج است. همچنین با افزایش تعداد داده‌های ورودی مدل، برنامه‌ریزی ژنتیک کند شده و گاهی قادر به ارائه نتایج نبوده درحالی‌که مدل برنامه‌ریزی بیان ژن این قابلیت را دارد که با تعداد ورودی‌ها و داده‌های بیشتر، نیز عمل مدل‌سازی را انجام دهد، نهایتاً مدل برنامه‌ریزی بیان ژن برای مدل‌سازی و پیش‌بینی جریان رودخانه قابلیت خوبی دارد. عباسی و همکاران (۱۳۹۹) کاربرد برنامه‌ریزی بیان ژن را در پیش‌بینی خشک‌سالی ایستگاه سینوپتیک تبریز موردبررسی قراردادند و دقت این مدل را قابل‌قبول اعلام کردند. نتایج پژوهش وانگ و همکاران در پیش‌بینی توان تبخیر و تعرق مرجع با GEP دقت زیاد این مدل را نشان می‌دهد. حافظ پرست و رحیمی (۱۳۹۹) به‌منظور ارزیابی تأثیر تغییر

1- Gene Expression Programming
2- M5 Model Tree

اقلیم بر رواناب منطقه سد جامیشان به بررسی و مقایسه‌ی مدل‌های SVM، GEP و IHACRES پرداختند. نتایج حاصل از پیش‌بینی دبی در هر سه مدل SVM، GEP و IHACRES حاکی از دقت بالاتر مدل‌های IHACRES و GEP نسبت به روش SVM است. یونسی و نوذری (۱۳۹۹) در پیش‌بینی کردن خشک‌سالی کوتاه‌مدت از مدل‌های تلفیقی شبکه عصبی مصنوعی-موجک و برنامه‌ریزی بیان - ژن-موجک استفاده کردند. نتایج پژوهش ایشان نشان داد از میان مدل‌های بررسی‌شده، برنامه‌ریزی بیان‌ژن-موجک با دقت بیش‌تری شاخص بارش معیار و وضعیت خشک‌سالی کوتاه‌مدت را پیش‌بینی می‌کند. همان‌طور که گفته شد، در تحقیقات گذشته پژوهشگران، اثرگذاری پیش‌پردازش (حذف خصوصیات ذاتی داده) به همراه پارامترهای مختلف مدل به‌صورت هم‌زمان بررسی نشده است. لذا هدف این تحقیق بررسی هم‌زمان خصوصیات داده و پارامترهای مهم در برنامه‌ریزی بیان ژن جهت پیش‌بینی‌های با دقت بیشتر در مراحل آموزش و صحت‌سنجی است.

مواد و روش‌ها

برنامه‌ریزی بیان ژن از جدیدترین الگوریتم‌های فرا کاوشی است که به دلیل دارا بودن دقت کافی، موردتوجه پژوهشگران قرارگرفته است. در این مدل در ابتدای فرآیند هیچ‌گونه رابطه تابعی در نظر گرفته نشده و این روش قادر به بهینه‌سازی ساختار مدل و مؤلفه‌های آن است. به‌طور خلاصه می‌توان گفت درGEP بهسازی‌ها در یک ساختار خطی اتفاق افتاده و سپس به‌صورت ساختار درختی بیان می‌شود و این موجب می‌شود تنها ژنوم اصلاح‌شده به نسل بعد منتقل‌شده و نیازی به ساختارهای سنگین برای تکثیر و جهش وجود نداشته باشد (Ferreia., 2006). به‌طورکلی مراحل اصلی الگوریتم‌های ذکرشده را می‌توان این‌گونه ذکر کرد که ابتدا فرآیند با تولید تصادفی کروموزوم‌ها از تعداد معینی افراد (جمعیت اولیه) آغاز می‌شود. سپس درGEP وGP این کروموزوم‌ها به‌صورت بیان درختی(ET) اظهار می‌شوند. در مرحله بعد با استفاده از تابع هدف میزان برازندگی هر فرد ارزیابی‌شده و بر اساس میزان عملکردشان انتخاب می‌شوند تا اصلاح‌شده و فرزندانی با ویژگی‌های جدید تشکیل شوند. فرزندان تولیدشده، دوباره تحت فرآیند توسعه‌ای قرارگرفته تا راه‌حل خوب و مناسبی پیدا شود(Ferreia., 2006). تغییرات ژنتیکی به‌واسطه عملگرهایی از قبیل جهش، وارون‌سازی، ترانهشIS ترانهشRIS، ترانهش ژن، ترکیب تک نقطه‌ای، ترکیب دونقطه‌ای و ترکیب ژنی است، انجام می‌شود (Ferreia., 2001). روشی کهGEP برای انتخاب افراد استفاده می‌کند به روش چرخ رولت[1] معروف است در این روش اعضا بر اساس میزان سازگاری نسبی آن‌ها انتخاب

1- Roulette-wheel

می‌شود. GEP برخلاف GA وGP، چندین عملگر ژنتیکی را برای تکثیر افراد با اصلاحات دارد (Lopes and Weinert., 2004). در GEP، هر ژن به‌صورت بیان درختی کدگذاری می‌شود. در مورد کروموزوم‌های چندژنی، تمامی بیان درختی‌ها با استفاده از تابع پیوند[2]، از محل گره ریشه[3] خود به یکدیگر متصل می‌شوند. (Ferreia., 2001).

یک کروموزوم از ژن‌ها تشکیل‌شده و معمولاً شامل بیش از یک ژن (کروموزوم چند ژن‌ها) است. هر ژن به یکسر[4] و یک دنباله[5] تقسیم می‌شود. اندازه رأس (h) به‌وسیله کاربر تعیین می‌شود، اما اندازه دنباله (t)، به‌صورت تابعی از h و پارامترn، به دست می‌آید. پارامترn، بیشترین تعداد پارامتر مستقل (arity) یافت شده در مجموعه توابع مورداستفاده در اجراست. رابطه (۱)، طول دنباله را با توجه به سایر پارامترها به دست می‌دهد (Lopes and Weinert., 2004).

$$t = h(n-1)+1 \quad (۱)$$

به‌عنوان‌مثال یک ژن را در نظر بگیرید که از { Q, *, /, -, +, a, b } تشکیل‌شده است. در این مورد تعداد متغیرهای مستقل n=۲ است. برای مقادیر طول سر h =۱۰ و طول دنباله t =۱۱ ، طول ژن برابر با ۲۱ خواهد بود (Ferreia., 2001).

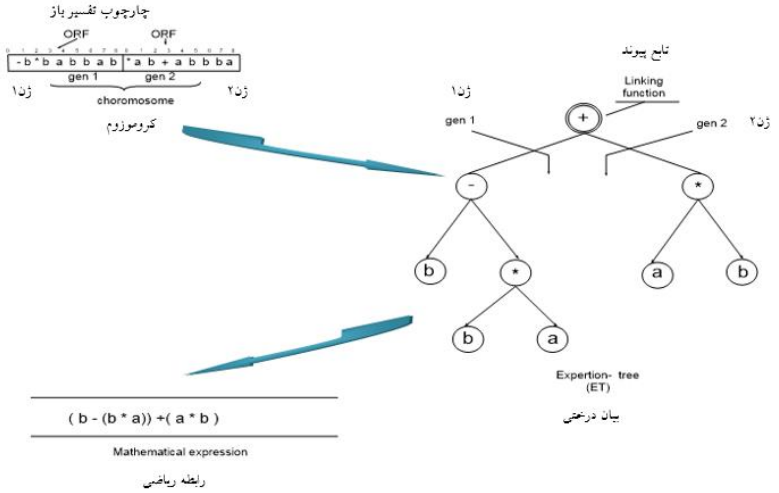
مسائل رگرسیونی نمادین، با استفاده از مجموعه‌ای از توابع و مجموعه‌ای از ترمینال‌ها، مدل‌سازی می‌شوند. مجموعه توابع، معمولاً شامل توابع اصلی حسابی {/، *، −، +}، توابع مثلثاتی یا هر نوع تابع ریاضی دیگر {x2, exp, log, cos, …} و یا توابع تعریف‌شده توسط کاربر است. توابع و ترمینال‌ها، در بخش سر ژن وجود دارند و در قسمت دنباله، فقط ترمینال‌ها وجود دارند (Lopes and Weinert., 2004).

یکی از موارد مهم در GEP، تعیین شاخص اعتبارسنجی تابع هدف است و هدف آن، یافتن راه‌حلی است که برای تمامی موارد برازش به‌اندازه یک خطای معین به‌خوبی عمل کند. برخی از توابع برازش مورداستفاده جذر میانگین مربعات خطا (RMSE)، میانگین خطای مطلق (MAE) و ضریب همبستگی (R) است (Lopes et al., 2004). تابع برازش نمایشگر کیفیت یک کروموزوم به‌عنوان یک جواب از مسئله است و میزان شایستگی یک کروموزوم را در بین جمعیت کروموزومی نشان می‌دهد.

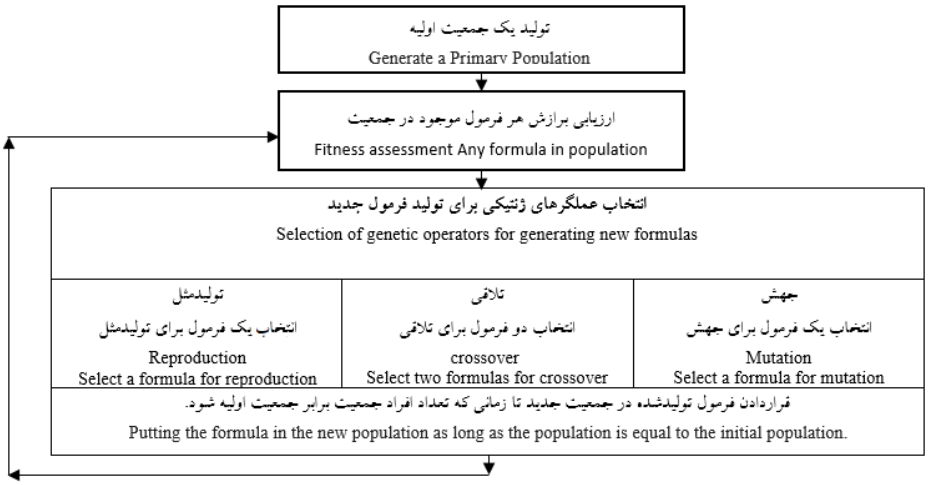
فرایند GEP مانند سایر الگوریتم‌ها زمانی خاتمه می‌یابد که به ضابطه تعیین‌شده‌ای مثل میزان معین خطای میانگین رسیده باشد و یا چرخه فرآیند به تعداد دلخواهی ادامه یافته، در این صورت بهترین راه‌حل یافت‌شده تابه‌حال گزارش داده‌شده و در غیر این صورت

2- Linking Function
3- Root Node
4- Head
5- Tail

بهترین راه‌حل از نسل حاضر نگه داشته‌می‌شود و بقیه راه‌حل‌ها به فرآیندی گزینشی واگذار می‌شوند که بر اساس آن بهترین افراد شانس بیشتری برای تولید فرزندان دارند.



شکل۱- کدگذاری به فرم رشته خطی و کد برداری به‌صورت ET یک کروموزوم با دو ژن در GEP



شکل ۲- فلوچارت برنامه‌ریزی بیان ژن(Nouryazdan(2015

با تکرار روند ذکرشده و با پیشرفته نسل به جلو کیفیت جمعیت بهبودیافته و به جواب بهینه نزدیک می‌شود(Lopes and Weinert., 2004). ازلحاظ ریاضی، برازش f_i از یک برنامه انفرادی i به‌صورت رابطه(۲) بیان می‌شود.

$$f_i = \sum_{j=1}^{n} \left(R - \left| \frac{P_{ij} - T_j}{T_j} .100 \right| \right) \qquad (۲)$$

که در آن R طول محدوده انتخابی، P_{ij} مقدار پیش‌بینی‌شده به‌وسیله برنامه انفرادی i برای مورد برازش j (از میان n مورد برازش) T_j مقدار هدف برای مورد برازش j است. قابل‌توجه است که عبارت داخل قدر مطلق متناظر با درصد خطای نسبی است (Ferreia., 2006). در این تحقیق طول محدوده انتخابی برابر با ۱۰۰ (R=100)، مجموعه‌ای شامل ۱۰ برازش (n=10) انتخاب گردیده است، بنابراین مقدار تابع برازش برابر با ۱۰۰۰ خواهد بود (f_{max} =1000) مزیت این نوع تابع برازش این است که سیستم می‌تواند با استفاده از آن، راه‌حل بهینه را پیدا کند. جذر میانگین مربعات خطا (RMSE)[1]، به‌عنوان معیار خطای تابع برازش انتخاب‌شده است. کروموزوم‌ها در GEP، معمولاً از چند ژن با طول مساوی تشکیل‌شده‌اند. برای هر مسئله یا هر اجرا، تعداد ژن‌ها نیز همانند طول رأس قابل انتخاب است. در ادامه در شکل ۲ فلوچارت مدل برنامه‌ریزی بیان ژن و مراحل مختلف الگوریتم آن ارائه‌شده است.

1- Root Mean Squared Error (RMSE)

بررسی توابع خودهمبستگی(ACF)

تابع خودهمبستگی یک روش برای بیان وابستگی زمانی در ساختار یک سری زمانی است.

$$\rho_k=\frac{\sum_{i=1}^{n-k}(z_i-\bar{z})(z_{i+k}-\bar{z})}{\sum_{i=1}^{n}(z_i-\bar{z})^2} \qquad -1 \le \rho_k \le +1 \qquad (۳)$$

ρ_k :مقدار تابع خودهمبستگی سری زمانی با تأخیر k

z_i و z_{i+k} : مقادیر متغیرها یا داده‌های سری زمانی در مرحله زمانی i و مرحله با تأخیر زمانی k، $\bar{z}$: مقدار میانگین مربوط به متغیرها (جعفر زاده و همکاران،۱۳۹۲).

معیارهای ارزیابی

در این تحقیق برای ارزیابی توانایی و دقت مدل برنامه‌ریزی بیان ژن از نمایه‌های ضریب همبستگی (R) و میانگین مربعات خطا (RMSE) و ضریب نش_ساتکلیف (NASH) استفاده شد،که به ترتیب با استفاده از روابط زیر قابل‌محاسبه است:

$$R=\frac{\sum_{i=1}^{n}(Qi-\overline{Qi})(\widehat{Qi}-\overline{\widehat{Qi}})}{\sqrt{\sum_{i=1}^{n}(Qi-\overline{Qi})^2\sum_{i=1}^{n}(\widehat{Qi}-\overline{\widehat{Qi}})^2}} \qquad (۴)$$

$$RMS=\sqrt{\frac{\sum_{i=1}^{n}(Qi-\widehat{Qi})^2}{n}} \qquad (۵)$$

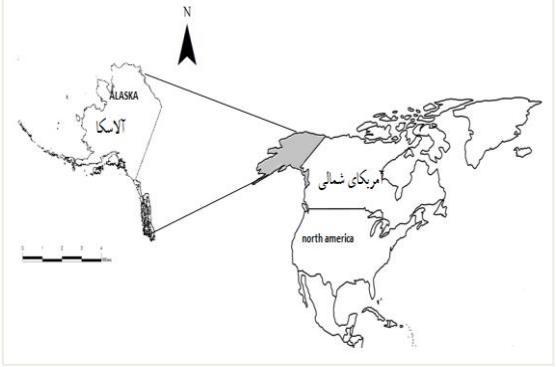
$$NASH=1-\frac{\sum_{t=1}^{T}(\hat{Q}_i^t-Q_i^t)^2}{\sum_{t=1}^{T}(Q_i^t-\overline{Q_i})^2} \qquad (۶)$$

$$NRMSE=\frac{RMSE*100}{\overline{Qi}} \qquad (۷)$$

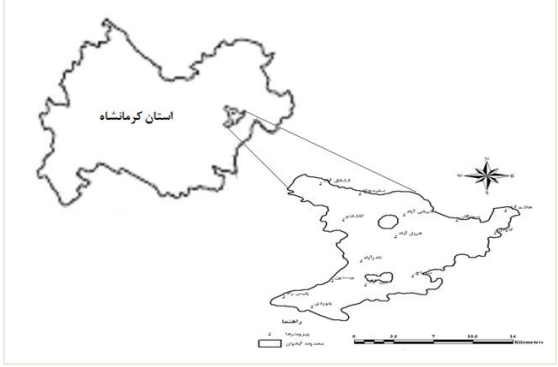
که در آن Qi مقادیر مشاهده‌شده، $\widehat{Qi}$ مقادیر محاسبه‌شده، $\overline{Qi}$ میانگین مقادیر مشاهده‌شده و $\overline{\widehat{Qi}}$ میانگین مقادیر محاسبه‌شده است. علاوه بر معیارهای فوق از نمودارهای مشاهداتی محاسباتی نیز جهت ارزیابی مدل‌ها استفاده گردید. در این تحقیق برای مشاهده وجود روند یا تناوب و نرمال بودن داده‌ها از نرم‌افزار Minitab استفاده شد. همچنین آزمون نرمال بودن داده‌ها از رسم گرافیکی داده‌ها در نمودار توزیع نرمال و با استفاده از آزمون اندرسون‌دارلینگ و در سطح اطمینان ۹۵ درصد صورت گرفت. برای به‌کارگیری روش برنامه‌ریزی بیان ژن از نرم‌افزار Genexprotools5.0 و تنظیمات پیش‌فرض برنامه استفاده شد.

منطقه مطالعاتی

برای بررسی اثرگذاری هم‌زمان خصوصیات سری زمانی و پارامترهای تأثیرگذار مدل بر نتایج، از مقادیر اندازه‌گیری شده دمای ماهانه در منطقه آلاسکا واقع در شمال غربی آمریکای شمالی با طول دوره ۵۰ سال و دارای اقلیم سرد و خشک استفاده گردید. تعداد کل داده‌ها در طول دوره (۲۰۱۷-۱۹۶۸) معادل ۶۰۰ داده است. سری زمانی دیگر، داده‌های عمق آب زیرزمینی دشت چمچمال واقع در استان کرمانشاه با طول دوره آماری۱۲سال و دارای اقلیم کوهستانی است. از ۷۰ درصد ابتدایی داده‌های مذکور به‌عنوان آموزش و از۳۰ درصد باقیمانده جهت صحت‌سنجی استفاده گردید.



شکل۳ - موقعیت جغرافیایی آلاسکا



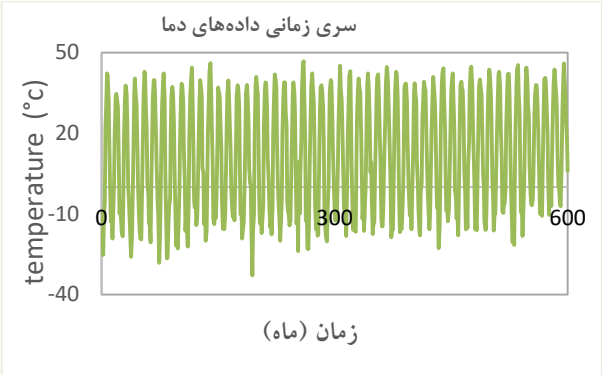
شکل٤ - موقعیت جغرافیایی ایستگاه چمچمال

بحث و نتایج

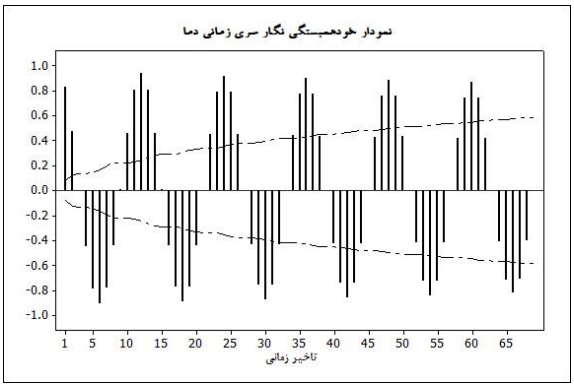
جهت بررسی ایستایی سری زمانی دما، نمودارهای آنالیز حاصل در شکل ۵ (شکل‌های الف تا د) ارائه‌شده است. همان‌طور که از اشکال ذکرشده مشخص است، داده‌های دما نرمال نیستند، تناوب دارند و با توجه به شیب‌خط برازش داده‌شده در نمودار روند این داده‌ها روند ندارند.

با توجه به شکل (ب) نمودار ACF سری زمانی دمای آلاسکا نشان‌دهنده وجود تناوب سینوسی در این سری زمانی است. این تناوب نامیرا است و طی تأخیرهای مختلف همچنان در سری زمانی وجود دارد. همچنین با توجه به نمودار ACF قدر مطلق بیشترین مقدار تابع

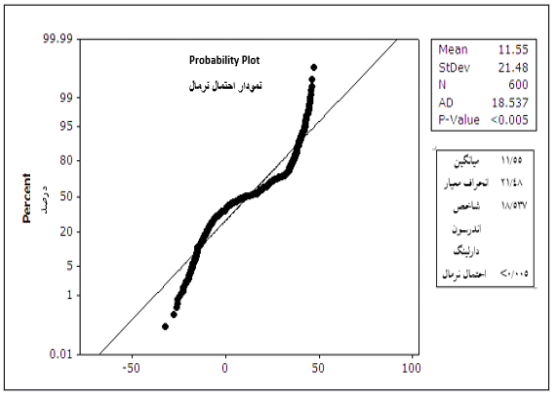
خودهمبستگی سری زمانی دمای آلاسکا نزدیک به یک است و این امر نشان‌دهنده وابستگی زمانی داده‌های سری زمانی است با توجه به شکل (ج)، مقادیر P-Value برای سری زمانی دمای ماهانه کمتر از ۰/۰۵ است. بنابراین این سری زمانی در سطح احتمال ۹۵درصد و با آزمون اندرسون دارلینگ نرمال نیست. با توجه به شکل (د) ناچیز بودن مقدار ضریب t نشان می‌دهد که سری زمانی ماهانه دمای ماهانه فاقد روند است.



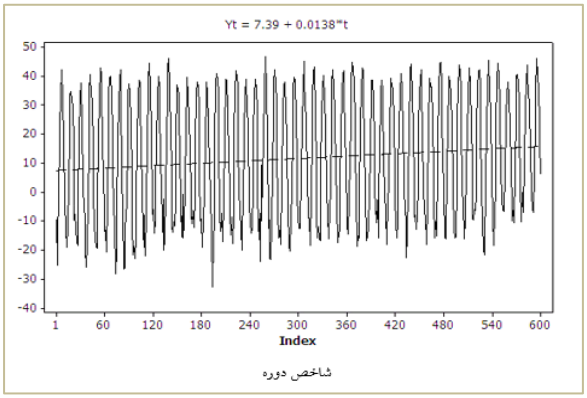
(الف)



(ب)



(ج)



(د)

شکل۵-نمودارهای سری زمانی داده‌های دما، (الف) نمودار سری زمانی داده‌های دما،(ب) نمودار ACF سری زمانی داده‌های دما، (ج) نمودار نرمال سری زمانی داده‌های دما ، (د) نمودار روند سری زمانی داده‌های دما

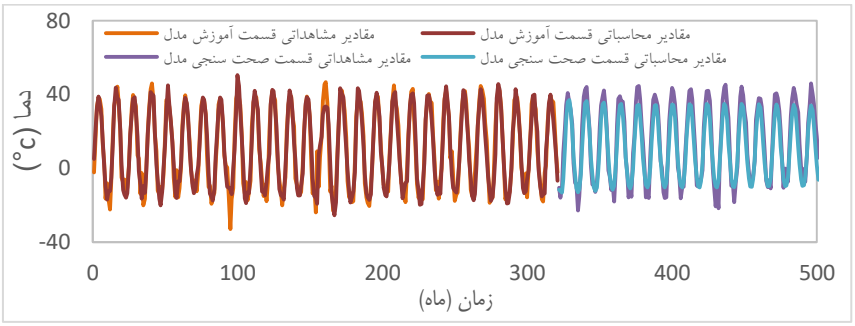
برای اجرای نرم‌افزار بیان ژن پارامترهای مربوط به تعیین عملگرهای ژنتیک و توابع برازش و معیار برازندگی تابع ثابت در نظر گرفته شد و از تنظیمات پیش‌فرض مدل که در جدول ۱ ارائه‌شده، استفاده شد. برای بررسی تأثیر پارامترهای خود مدل بر نتایج نهایی نرم‌افزار، ساختار کروموزوم‌ها، Embedding Dimension (تعداد تأخیر) و Head Size (اندازه رأس) تغییر داده‌شده. برای هر پارامتر سه مقدار مختلف در نظر گرفته شد که نتایج اجرای مدل برای هر پارامتر ارائه‌شده است. در ابتدا از داده‌های ماهانه دمای آلاسکا با دوره زمانی ۵۰ ساله اجرا گرفته شد. تعداد داده‌ها ۶۰۰ ماه است که از ۷۰ درصد داده‌ها برای آموزش مدل و از ۳۰ درصد باقیمانده برای صحت-سنجی مدل استفاده شد. نتایج حاصل در جداول ۲ و ۳ و۴ ارائه‌شده است.

جدول۱- تنظیمات عمومی نرم‌افزار

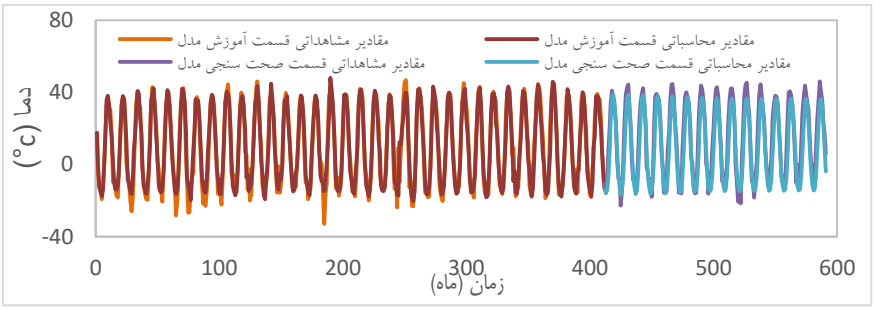
تعداد ژن‌ها در هر کروموزوم	تابع پیوند	نوع تابع برازش	توابع مورداستفاده
۵	Avg2	RMSE	توابع پیش‌فرض سری زمانی

جدول۲- نتایج بیان ژن برای سری زمانی ۵۰ ساله دمای ماهانه آلاسکا با تعداد تأخیر متغیر در برنامه‌ریزی بیان ژن

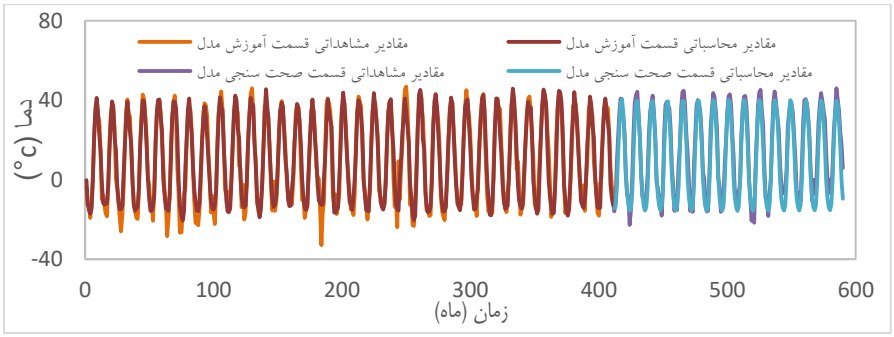
شماره اجرا	تعداد تأخیرها	training				testing			
		R	RMSE	NRMSE	NASH	R	RMSE	NASH	NRMSE
اجرا۱	۳	۰/۹۷	۸/۲	۷۳/۶	۰/۹۷	۰/۹۵	۶/۷	۰/۹۰	۴۱/۵
اجرا۲	۱۰	۰/۹۸	۶	۵۲/۰۶	۰/۹۶	۰/۹۷	۵/۵۴	۰/۹۳	۳۸/۳
اجرا۳	۲۵	۰/۹۷	۵/۸	۵۲/۰۶	۰/۹۵	۰/۹۷	۴/۹	۰/۹۴	۳۳/۹



الف



ب



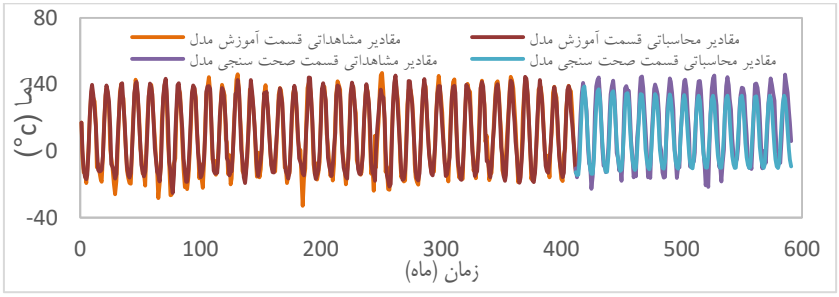
ج

شکل ٦- الف، ب و ج نمودار مشاهداتی محاسباتی دمای ماهانه آلاسکا به ترتیب با تعداد تأخیر ۳، ۱۰ و ۲۵

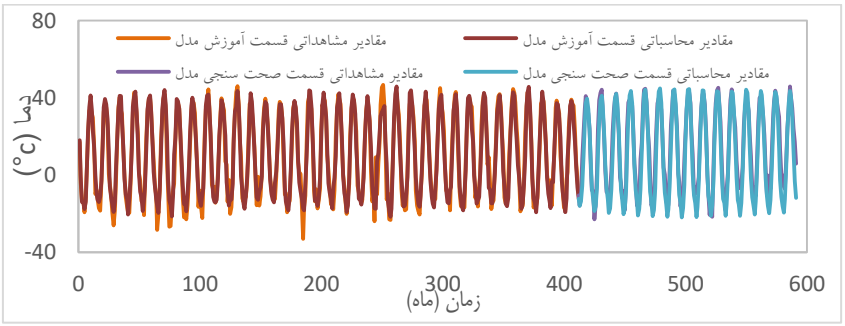
نرم‌افزار در زمان کوتاهی نتایج فوق را حاصل نمود. همان‌طور که از نتایج شکل ۶ پیداست مدل توانسته به‌خوبی روند تغییرات سری زمانی دما را در هر اجرا پیش‌بینی کند. شاخص R در هردو قسمت آموزش و صحت‌سنجی مدل برای اجراهای مختلف، در محدوده ۰/۹۵-۰/۹۷است و تغییر پارامتر تعداد تأخیرها اثر چشم‌گیری بر نتایج نداشته است. برای بررسی‌های بیشتر، برای پارامتر اندازه رأس هم همانند مرحله قبل سه مقدار مختلف در نظر گرفته شد و از نرم‌افزار اجرا گرفته شد. نتایج این بخش در جداول ۳ و ۴ و نمودارهای حاصل‌شده در اشکال ۹ و۱۰ نشان داده‌شده است.

جدول ۳- نتایج بیان ژن برای سری زمانی ۵۰ ساله دمای ماهانه آلاسکا با تعداد اندازه رأس متغیر در برنامه‌ریزی بیان ژن

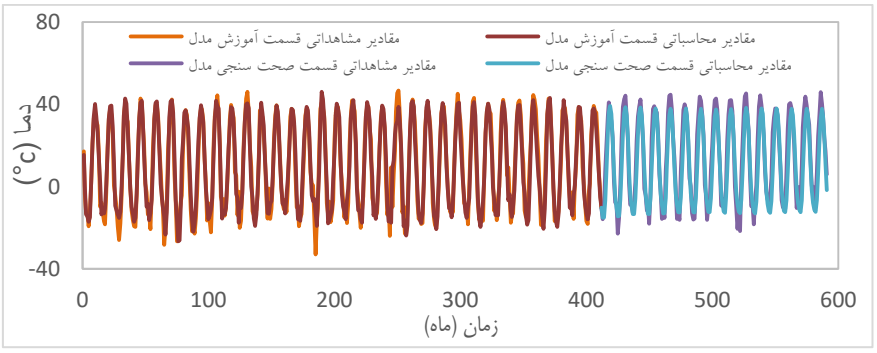
شماره اجرا	اندازه رأس	training				testing			
		R	RMSE	NASH	NRMSE	R	RMSE	NASH	NRMSE
اجرا۱	۳	۰/۹۵	۱۰	۰/۹۵	۸۹/۷	۰/۹۶	۵/۹	۰/۹۱	۴۰/۸
اجرا۲	۸	۰/۹۲	۵/۳۸	۰/۹۵	۴۸/۲۹	۰/۹۶	۶/۰۲	۰/۹۴	۴۱/۶
اجرا۳	۱۵	۰/۹۷	۷/۵	۰/۹۵	۶۷/۳	۰/۹۶	۶/۰۸	۰/۹۲	۴۲/۰



الف



ب



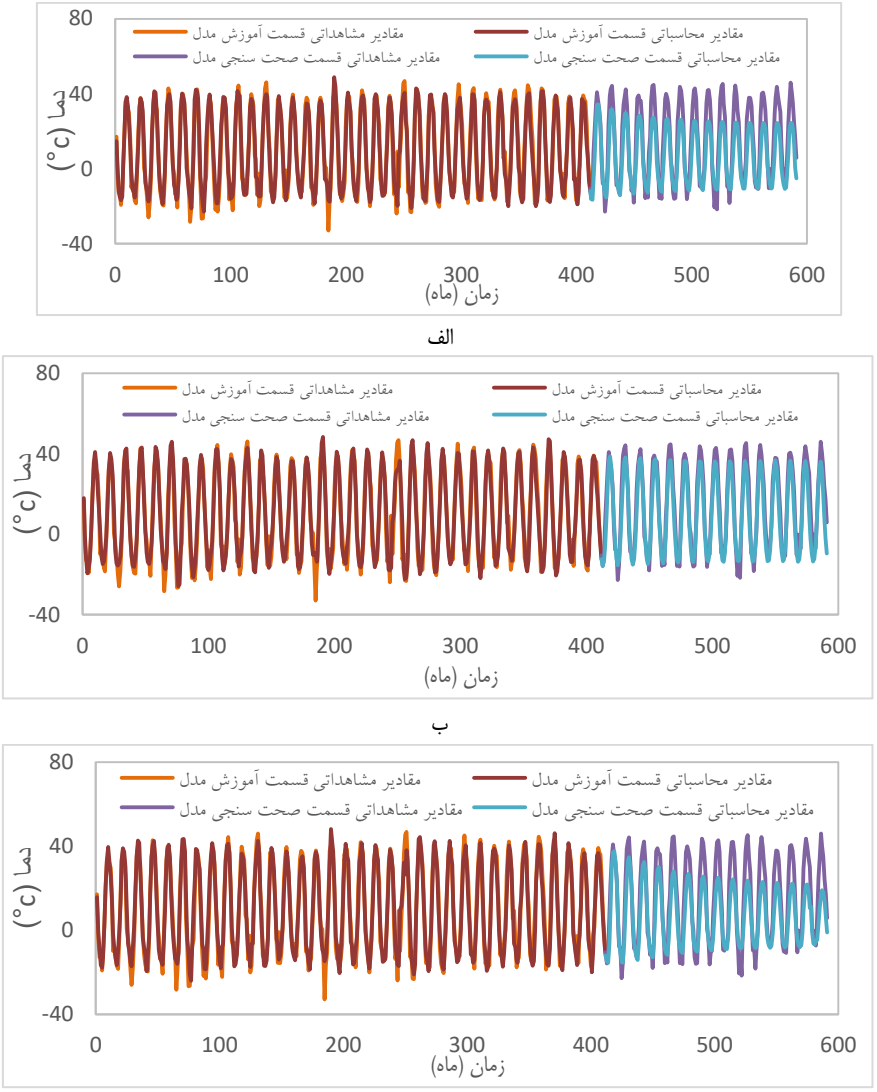
ج

شکل۹- الف، ب و ج نمودار مشاهداتی محاسباتی دمای ماهانه آلاسکا به ترتیب با تعداد اندازه رأس ۳، ۸ و ۱۵

همان‌طور که جدول (۳) نشان می‌دهد، با تعداد رأس متغیر هم نتایج مدل‌سازی توسط برنامه‌ریزی بیان ژن دارای معیارهای آماری قابل‌قبول در هردو قسمت آموزش و صحت‌سنجی است. به‌طوری‌که در هردو قسمت آموزش و صحت‌سنجی ضرایب به‌دست‌آمده برای معیارهای همبستگی و نش-ساتکلیف بیشتر از ۰/۹۰ هستند.

جدول ٤- نتایج بیان ژن برای سری زمانی ۵۰ ساله دمای ماهانه آلاسکا با تعداد کروموزوم متغیر در برنامه‌ریزی بیان ژن

شماره اجرا	تعداد کروموزوم	training				testing			
		R	RMSE	NRMSE	NASH	R	RMSE	NASH	NRMSE
اجرا۱	۳۰	۰/۹۷	۵/۵	۴۹/۴	۰/۹۹	۰/۹۶	۸/۶	۰/۹۳	۴۷/۷
اجرا۲	۲۵۰	۰/۹۷	۵/۴	۵۰/۳	۰/۹۹	۰/۹۷	۶/۹	۰/۹۳	۴۷/۷
اجرا۳	۴۰۰	۰/۹۶	۵/۶	۵۰/۳	۰/۹۹	۰/۹۴	۸/۰۴	۰/۹۱	۴۵/۶



الف

ب

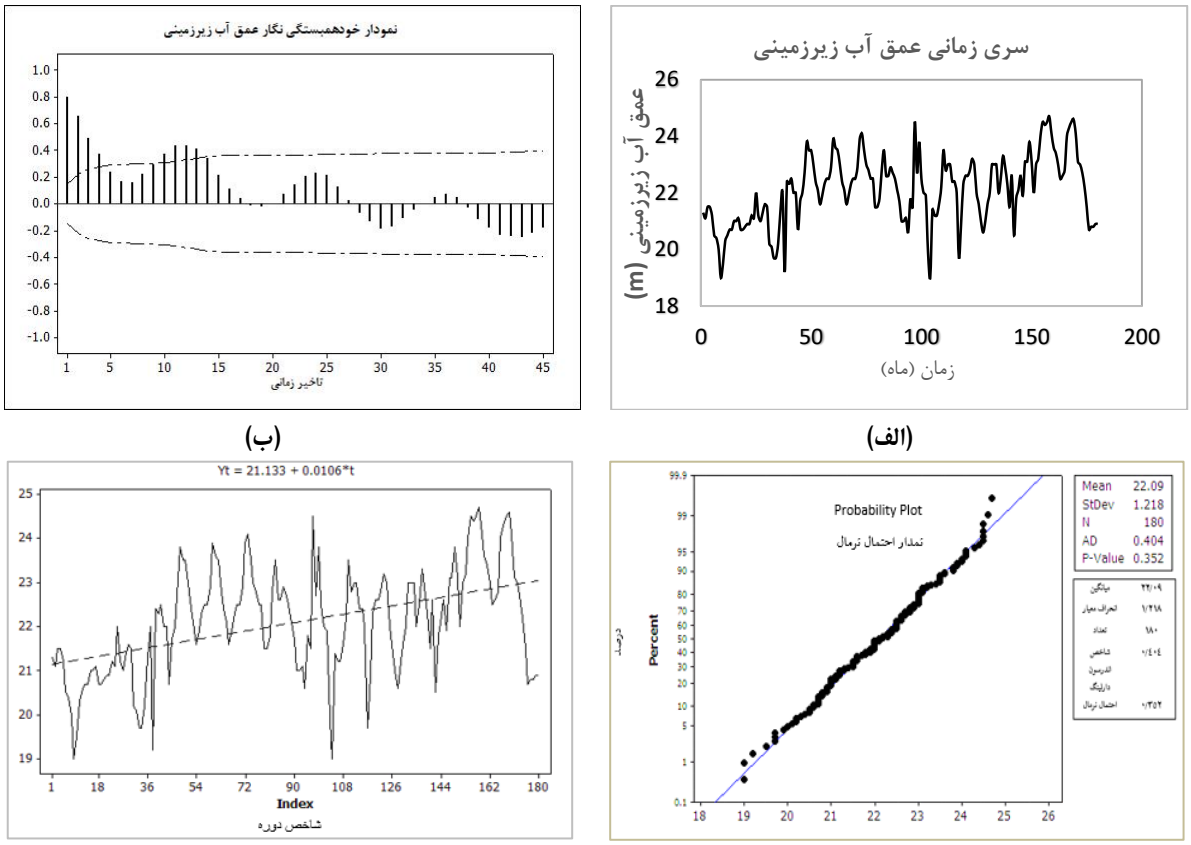
ج

شکل ۱۰- الف، ب و ج نمودار مشاهداتی محاسباتی دمای ماهانه آلاسکا به ترتیب با تعداد کروموزوم ۳۰، ۲۵۰ و ۴۰۰

در این مرحله نیز همانند مرحله قبل، نتایج قابل قبولی حاصل‌شده است و نمودارهای مقادیر مشاهده‌شده و محاسبه‌شده در هردو قسمت آموزش و صحت‌سنجی برهم منطبق هستند. برای بررسی‌های بیشتر، از یک سری زمانی دیگر که مربوط به عمق آب زیرزمینی است، استفاده شد. نمودارهای آنالیز آماری مربوط به این سری زمانی در شکل ۱۱ (نمودارهای الف-د) نشان داده‌شده است.

همان‌طور که شکل (ب) نشان می‌دهد، نمودار سری زمانی داده‌های سطح آب زیرزمینی چمچمال فاقد تناوب سینوسی است و مقادیر حداکثر ناپایداری در نمودار ACF این سری زمانی با تعداد تأخیرهای متفاوت کمتر می‌شود؛ بنابراین نمودارACF آن نزولی است.

داده‌های عمق آب زیرزمینی در سطح ۹۵درصد اطمینان نرمال هستند، تناوب ندارند و با توجه به شیب‌خط برازش داده‌شده در نمودار روند ندارند. همچنین با توجه به نمودار خودهمبستگی سری زمانی باگذشت زمان میرا می‌شود. پس می‌توان نتیجه گرفت که سری زمانی ایستا است. انتظار می‌رود که نرم‌افزار برنامه‌ریزی بیان ژن برای داده‌های عمق آب زیرزمینی با توجه به ایستا بودن سری زمانی و با توجه به معیارهای آماری R و RMSE نتایج قابل‌قبول‌تری ارائه دهد. در ادامه این فرضیه بررسی می‌شود. در مرحله دوم داده‌های عمق آب زیرزمینی وارد مدل شد و همانند سری زمانی دما از مدل GEP اجرا گرفته شد. ابتدا تعداد Embedding Dimension (تعداد تأخیر) متغیر در نظر گرفته شد و از سایر تنظیمات پیش‌فرض مدل استفاده شد. در ادامه Head Size و تعداد کروموزوم‌ها متغیر در نظر گرفته شد و از سایر تنظیمات پیش‌فرض مدل استفاده شد. نتایج در جدول‌های شماره ۵ و ۶ و ۷ ارائه شده است.



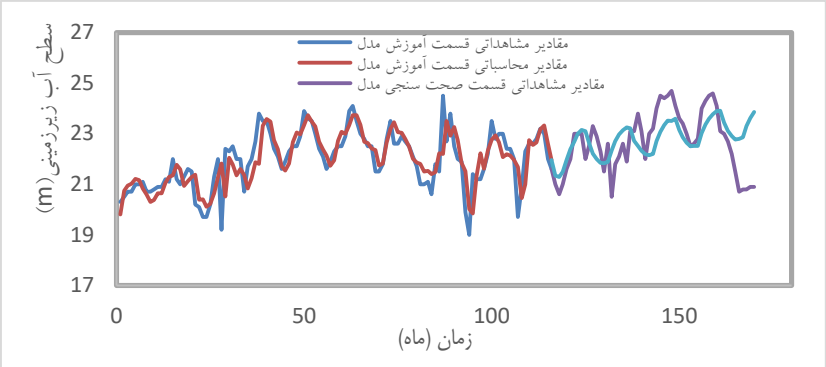
شکل ۱۱- نمودارهای سری زمانی عمق آب زیرزمینی، (الف) نمودار سری زمانی داده‌های آب زیرزمینی،(ب) نمودار ACF سری زمانی داده‌های آب زیرزمینی، (ج) نمودار نرمال سری زمانی داده‌های آب زیرزمینی، (د) نمودار روند سری زمانی داده‌های آب زیرزمینی

جدول ۵- نتایج بیان ژن برای سری زمانی سطح آب زیرزمینی چمچمال با تعداد تأخیر متغیر در برنامه‌ریزی بیان ژن

شماره اجرا	تعداد تأخیرها	training				Testing			
		R	RMSE	NASH	NRMSE	R	RMSE	NASH	NRMSE
اجرا۱	۳	۰/۷۷	۰/۷۳	۰/۷۰	۳/۳	-۰/۴	۱/۹	-۰/۱	**۸/٤**
اجرا۲	۱۰	۰/۷۸	۰/۶۸	۰/۷۲	۳/۱	۰/۴۴	۱/۰۷	۰/۱۵	**٤/۷**
اجرا۳	۲۵	۰/۷۳	۰/۷۶	۰/۶۵	۳/۴	۰/۱۴	۱/۵۱	۰/۰۲	**٦/۷**

همان‌طور که نتایج جدول فوق نشان می‌دهد، مدل برنامه‌ریزی بیان ژن در مدل‌سازی سطح آب زیرزمینی در قسمت آموزش مدل عملکرد بهتری نسبت به قسمت صحت‌سنجی داشته به‌طوری‌که ضریب همبستگی R در محدوده ۰/۷۸-۰/۷۳ حاصل‌شده است. اما مقدار ضریب همبستگی در قسمت صحت‌سنجی در بیشترین حالت برابر با ۰/۴۴ است؛ بنابراین مدل در پیش‌بینی سطح آب زیرزمینی عملکرد مناسبی ندارد و تغییر در هر پارامتر بیان ژن موجب حاصل شدن ضریب همبستگی بیشتر از ۰/۴۴ در قسمت صحت‌سنجی نشده است. نمودار اجرای شماره۲ مربوط به قسمت اثرگذاری پارامتر Embedding dimension (تعداد تأخیر) برای هردو قسمت آموزش و صحت‌سنجی در شکل ۱۲ ارائه‌شده است.

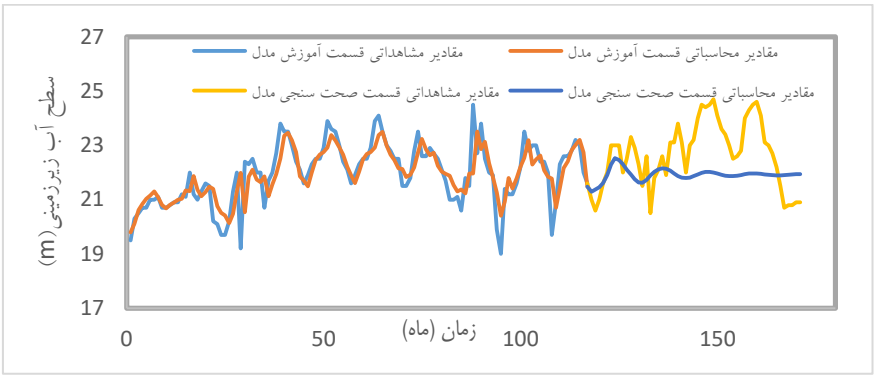
در ادامه تعداد رأس‌ها متغیر در نظر گرفته شد و نرم‌افزار اجرا شد. با توجه به جدول (۶) می‌توان اظهار داشت که تغییر در تعداد رأس هم موجب افزایش ضریب R بیشتر از ۰/۴۲ در قسمت صحت‌سنجی نشده است. در این مرحله هم همانند مرحله قبل عملکرد مدل در قسمت آموزش با حاصل شدن ضریب همبستگی برابر ۰/۸۱ بهتر از قسمت صحت‌سنجی است. در ادامه نمودار اجرای شماره ۳ در شکل ۱۳ ارائه‌شده است.



شکل۱۲- نمودار مشاهداتی محاسباتی سطح آب زیرزمینی چمچمال با تعداد تأخیر ۱۰

جدول ٦- نتایج بیان ژن برای سری زمانی سطح آب زیرزمینی چمچمال با تعداد اندازه رأس متغیر در برنامه‌ریزی بیان ژن

شماره اجرا	اندازه رأس	training				Testing			
		R	RMSE	NASH	NRMSE	R	RMSE	NASH	NRMSE
اجرا۱	۳	۰/۷۴	۰/۷۶	۰/۶۵	۳/۵	۰/۳۴	۱/۳۳	۰/۱	۵/۹
اجرا۲	۸	۰/۶۸	۰/۸۳	۰/۵۵	۳/۸	۰/۱۱	۱/۱۹	۰/۰۱	۵/۲
اجرا۳	۱۵	۰/۸۱	۰/۶۵	۰/۷۸	۲/۹	۰/۴۲	۱/۲۶	۰/۱۵	۵/۴



شکل۱۳- نمودار مشاهداتی محاسباتی سطح آب زیرزمینی چمچمال با تعداد اندازه رأس ۲۵

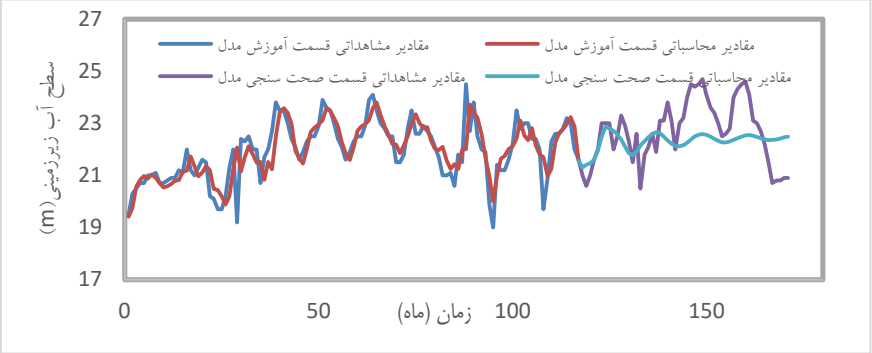
در ادامه تعداد کروموزوم‌ها متغیر در نظر گرفته شد و نرم‌افزار اجرا شد. این نتایج در جدول ۷ ارائه شده است.

جدول ۷- نتایج بیان ژن برای سری زمانی سطح آب زیرزمینی چمچمال با تعداد کروموزوم متغیر در برنامه‌ریزی بیان ژن

شماره اجرا	کروموزوم تعداد	Training				Testing			
		R	RMSE	NASH	NRMSE	R	RMSE	NASH	NRMSE
اجرا۱	۳۰	۰/۷۸	۰/۶۹	۰/۵۶	۳/۱	۰/۲۸	۱/۲۲	۰/۰۵	۵/۴
اجرا۲	۲۵۰	۰/۸۲	۰/۶۴	۰/۶۲	۲/۹	۰/۴۴	۱/۳۶	۰/۱۰	۶
اجرا۳	۴۰۰	۰/۸۰	۰/۶۶	۰/۶۰	۳	۰/۳۳	۱/۳	۰/۱۳	۵/۷

نمودارهای فوق نیز همانند نتایج حاصل‌شده برای سری زمانی سطح آب زیرزمینی چمچمال در مراحل قبل نیز بیانگر عدم توانایی مدل در تخمین مقادیر داده‌ها در قسمت صحت‌سنجی است. در این حالت با تغییر در هر پارامتر مدل، بهترین نتیجه در قسمت صحت-سنجی برای معیار R برابر با ۴۴ درصد حاصل شد. همان‌طور که در شکل‌های ۱۴-۱۲ پیداست، نرم‌افزار در مدل‌سازی و پیش‌بینی سری زمانی سطح آب زیرزمینی در قسمت صحت‌سنجی عملکرد قابل‌قبولی نداشته و ضعیف عملکرده است. اما در قسمت آموزش، مدل برنامه-ریزی بیان ژن دارای عملکرد بهتر و مناسب‌تری است. دلیل این امر به خاطر پیش‌پردازشی است که روی‌داده‌ها انجام‌شده است.

در ادامه همانند مراحل قبل با سناریوهای مختلف نرم‌افزار اجرا شد. ابتدا اندازه رأس متغیر در نظر گرفته شد سپس با تعداد کروموزوم‌های متغیر نرم‌افزار اجرا شد. نتایج هر بخش در جداول ۹ و ۱۰ ارائه شده است.



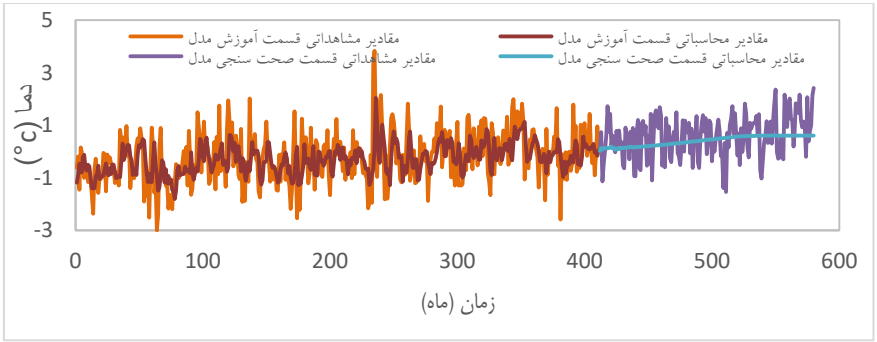
شکل ۱۴- نمودار مشاهداتی محاسباتی سطح آب زیرزمینی چمچمال با تعداد کروموزوم ۲۵۰

لذا استفاده از روش‌های پیش‌پردازش داده‌ها باعث افزایش عملکرد مدل GEP شده است. سلگی و همکاران (۱۳۹۷) نیز در مدل‌سازی جریان ماهانه رودخانه گاماسیاب از مدل برنامه‌ریزی بیان ژن استفاده کردند. نتایج آن‌ها نشان داد که عملکرد مدل ترکیبی از مدل ساده بهتر بوده است. دلیل این امر به خاطر پیش‌پردازشی است که روی‌داده‌ها پیاده شده بود. لذا استفاده از روش‌های پیش‌پردازش داده‌ها باعث افزایش عملکرد مدل GEP شده است. برای پاسخ به فرضیه اثرگذاری خصوصیت داده‌ها بر روی نتایج، به‌منظور حذف تناوب سری زمانی دما عمل استانداردسازی بر روی سری زمانی صورت گرفت. درنتیجه سری زمانی دما نرمال شد و تناوب آن حذف شد و مجدداً این سری زمانی وارد نرم‌افزار شد و همانند مراحل قبل از نرم‌افزار اجرا گرفته شد. نتایج در جداول (۱۰-۸) ارائه‌شده است.

جدول ۸- نتایج بیان ژن برای سری زمانی استانداردشده ۵۰ ساله دمای آلاسکا با تعداد تأخیر متغیر در برنامه‌ریزی بیان ژن

شماره اجرا	تعداد تأخیرها	training			testing		
		R	RMSE	NASH	R	RMSE	NASH
اجرا ۱	۳	۰/۴۲	۰/۸۷	۰/۱۳	-۰/۱	۱/۱۳	-۱/۰۲
اجرا ۲	۱۰	۰/۵	۰/۸۵	۰/۱۸	-۰/۱	۱/۳۲	-۱/۸
اجرا ۳	۲۵	۰/۳۴	۰/۹۴	۰	۰/۲۶	۰/۷۸	۰/۰۲

همان‌طور که جدول فوق نشان می‌دهد، بعد از اعمال استانداردسازی سری زمانی دما نتایج مدل‌سازی کاهش چشم‌گیری داشته و معیارهای آماری قابل قبولی حاصل نشده است. به‌طوری‌که بهترین نتیجه ضریب همبستگی (R) برای قسمت آموزش مدل برابر ۰/۵ یا ۵۰٪ و برای قسمت صحت‌سنجی برابر ۰/۲۶ حاصل‌شده است. نمودار اجرای شماره ۳ در این بخش در شکل نشان داده‌شده است.



شکل ۱۵- نمودار مشاهداتی محاسباتی دمای ماهانه استانداردشده آلاسکا با تعداد تأخیر ۲۵

جدول۹- نتایج بیان ژن برای سری زمانی استانداردشده ۵۰ ساله دمای آلاسکا با تعداد اندازه رأس متغیر در برنامه‌ریزی بیان ژن

شماره اجرا		اندازه رأس	training		testing		
		R	RMSE	NASH	R	RMSE	NASH
اجرا ۱	۳	-۰/۱	۱/۵۲	-۲/۶	۰/۴۸	۰/۸۵	۰/۱۵
اجرا ۲	۸	۰/۱۱	۱/۲۸	-۱/۶۱	۰/۵۲	۰/۸۴	۰/۲۱
اجرا ۳	۱۵	۰/۱۱	۱/۲۷	-۱/۵	۰/۵	۰/۸۴	۰/۲

همان‌طور که جدول (۹) نشان می‌دهد، نتایج مدل‌سازی سری زمانی استانداردشده دمای آلاسکا به طرز معنی‌داری ازلحاظ معیارهای کاهش‌یافته است. به‌طوری‌که بیشترین مقدار R در قسمت آموزش مدل ۰/۱۱ است و در قسمت صحت‌سنجی برای معیار R مقدار بیشتر از ۰/۵۲ حاصل نشده است.

جدول۱۰- نتایج بیان ژن برای سری زمانی استانداردشده ۵۰ ساله دمای آلاسکا با تعداد کروموزوم متغیر در برنامه‌ریزی بیان ژن

شماره اجرا		تعداد کروموزوم‌ها	training		testing		
		R	RMSE	NASH	R	RMSE	NASH
اجرا ۱	۳۰	۰/۱۹	۰/۹۸	-۰/۰۷	-۰/۲۱	۰/۹۸	-۰/۵
اجرا ۲	۲۵۰	۰/۱۲	۰/۹۷	-۰/۰۲	-۰/۰۷	۱/۵	-۱/۱
اجرا ۳	۴۰۰	۰/۵	۰/۸۵	۰/۲	-۰/۱۲	۱/۲۷	-۱/۵

همان‌طور که جدول (۱۰) نشان می‌دهد، نتایج مدل‌سازی سری زمانی استانداردشده دمای آلاسکا به طرز معنی‌داری ازلحاظ معیارهای آماری کاهش‌یافته است. به‌طوری‌که بیشترین مقدار R در قسمت آموزش مدل ۰/۱۹ است و در قسمت صحت‌سنجی برای معیار R مقدار منفی به‌دست‌آمده است. نتایج جدول فوق نشان می‌دهد که نرم‌افزار قادر به مدل‌سازی سری زمانی استانداردشده دمای آلاسکا نبوده و معیارهای آماری قابل‌قبول حاصل نشده است. اعمال پیش-پردازش بر روی سری زمانی دما و حذف خصوصیت تناوبی آن، باعث کاهش معنی‌دار معیارهای آماری گردید و عملکرد مدل به طرز چشم-گیری کاهش یافت. موئینی و همکاران بر غالب بودن ترم تناوب و اثر آن در بالا بودن عملکرد نتایج پیش‌بینی مدل‌های ARIMA بر سری زمانی دما اشاره داشتند. آن‌ها در مطالعات خود از شش سری زمانی شامل ۴ سری جریان و ۲ سری دمای آب استفاده نمودند. برای تجزیه‌وتحلیل داده‌ها ۹۲۲۸ مدل معرفی اجرا گردیده و نتایج آن‌ها نشان داده است که استانداردسازی بهترین روش برای حذف تناوب سری زمانی دمای ماهانه آب و مهم‌ترین عامل در دقت مدل‌های پیش‌بینی است (Moeeni et al., 2017)

نتیجه‌گیری

در این تحقیق برای پاسخ به فرضیه تأثیر خصوصیات مربوط به یک سری زمانی بر روی نتایج مدل‌سازی، از الگوریتم برنامه‌ریزی بیان ژن و نرم‌افزارGeneXprotools استفاده گردید و از دونقطه نظر تأثیر پیش‌پردازش سری زمانی شامل تناوب، روند، و نرمال بودن و پارامترهای تأثیرگذار خود نرم‌افزار شامل Embedding Dimension (تعداد تأخیر)، Head Size (اندازه رأس)، و اندازه طول کروموزوم و تأثیر هم‌زمان آن‌ها در پیش‌بینی نتایج بهتر موردبررسی قرار گرفت. به همین منظور ابتدا از سری زمانی دمای ماهانه منطقه آلاسکا که دارای خصوصیت قطعی تناوب است، استفاده شد. با تغییر هر پارامتر تأثیرگذار بیان ژن بهترین نتایج حاصل‌شده و تغییر در پارامترهای مدل تأثیر معنی‌داری در نتایج نداشت. نتایج به‌دست‌آمده برای سری زمانی دما نشان‌دهنده عملکرد مناسب و قابل‌قبول و مورد تائید این الگوریتم ازلحاظ معیارهای آماری است. همان‌طور که در شکل‌های شماره (۱۰-۶) مشخص است، نمودارهای برازش داده‌شده توسط مدل بر نمودار خود داده‌ها منطبق است و نمودارهای مشاهداتی و محاسباتی دارای هم‌پوشانی خوبی باهم هستند. به‌طوری‌که مقدار ضریب همبستگی (R) در هر اجرا با هر پارامتر تأثیرگذار بیان ژن برای هردو قسمت آموزش و صحت‌سنجی در محدوده ۰/۹۰-۰/۹۷ است. علت این امر وجودترم تناوب که از خصوصیات داده‌های ماهانه است، است. برای اثبات این موضوع با استفاده از استانداردسازی این ویژگی از داده‌های ماهانه حذف شد و با تغییر در هر پارامتر مدل بیان ژن نتایج معیار R بهتر از ۵۲ درصد نشد. نمودارهای مشاهداتی و محاسباتی در شکل۱۵ نشان می‌دهد که مدل در این حالت قادر نبوده مقادیر را تخمین بزند. در مرحله بعد از داده‌های ایستگاه چمچمال که فاقد خصوصیت تناوبی بود و ایستا بود، استفاده شد. در این حالت با تغییر در هر پارامتر مدل، بهترین نتیجه در قسمت صحت‌سنجی برای معیار R برابر با ۴۴ درصد حاصل شد. همان‌طور که در شکل‌های

(۱۴-۱۲) پیداست، نرم‌افزار در مدل‌سازی سری عمق آب زیرزمینی، عملکرد قابل‌قبولی نداشته و ضعیف عملکرده است.

درنهایت می‌توان نتیجه گرفت تأثیر پیش‌پردازش سری زمانی در نتایج پیش‌بینی با مدل برنامه‌ریزی بیان ژن خیلی بیشتر از پارامترهای خود مدل است. دارا بودن مؤلفه تناوب برای سری زمانی موجب عملکرد عالی مدل می‌شود. همچنین در مورد نتایج به‌دست‌آمده برای سری زمانی عمق آب زیرزمینی می‌توان اظهار داشت که تغییر در پارامترهای خود نرم‌افزار تنها درصد کمی حدود ۱۰درصد در بهبود نتایج تأثیر دارد. و تنها خصوصیات خود داده‌ها است که در عملکرد مدل اثرگذار است.

منابع

امامقلی زاده، ص.، کریمی‌دمنه، ر. و مهدی پناه، ح. ۱۳۹۵. برآورد رواناب حوضه آبریز کسیلیان با استفاده از روش برنامه‌ریزی بیان ژن. مجله مطالعات منابع طبیعی،محیط‌زیست و کشاورزی. ۲ (۴ و۵): ۱-۷.

ثانی خانی، ه.، نیک پور، م.، فرسادی‌زاده، د. و معیری م.م. ۱۳۹۴. پیش‌بینی بار معلق رودخانه با استفاده از سامانه‌های هوشمند. مجله پژوهش آب ایران. ۹(۲):۱۶۵-۱۶۸.

جعفر زاده، ج.، رستم زاده، ه. و اسدی، ا. ۱۳۹۶. مدل‌سازی زمانی تراز آب زیرزمینی با استفاده از روش‌های پایه تحلیل سری‌های زمانی (مطالعه موردی: دشت اردبیل). نشریه دانش آب‌وخاک. ۲۷(۴): ۱۹۶-۱۸۵ .

حافظ پرست مودت،م. و رحیمی، ب. ۱۳۹۹. مقایسه مدل‌های SVM ، GEP و IHACRES در پیش‌بینی تغییرات رواناب ناشی از تغییر اقلیم (مطالعه موردی: سد جامیشان). نشریه تحقیقات آب‌وخاک ایران. ۵۹(۱۰): ۲۴۸۳-۲۴۹۹.

خادم پور، ف. خاشعی سیوکی، ع. و امیرآبادی زاده، م. ۱۳۹۷. بررسی عملکرد روش برنامه‌ریزی بیان ژن در پیش‌بینی تابش خورشیدی روزانه در گستره ایران. نشریه پژوهش‌های اقلیم‌شناسی. ۹(۳۶): ۵۶-۴۳.

زمانی، ر.، احمدی، ف. و رادمنش، ف.، ۱۳۹۳. مقایسه روش‌های برنامه‌ریزی بیان ژن، سری زمانی غیرخطی، خطی و شبکه عصبی مصنوعی در تخمین دبی روزانه (مطالعه موردی: رودخانه کارون)، نشریه آب‌وخاک (علوم و صنایع کشاورزی). ۲۸ (۶) : ۱۱۸۲-۱۱۷۲.

سلگی، ا.، زارعی، ح.، دارابی، م. و ده کهنه، ص. ۱۳۹۷. پیش‌بینی بارش ماهانه با استفاده از مدل‌های برنامه‌ریزی بیان ژن و ماشین بردار پشتیبان، نشریه تحقیقات کاربردی علوم جغرافیایی. ۱۸ (۵۰):۰۹۱-۱۰۳.

ظهیری، ع. و قربانی، خ. ۱۳۹۲. شبیه‌سازی دبی جریان در مقاطع مرکب به کمک مدل درخت تصمیمM5 . نشریه پژوهش‌های حفاظت آب‌وخاک.۲۰ (۳):۱۱۳-۱۳۲.

عباسی، ع.، خلیلی،ک.، بهمنش، ج. و شیرزاد، ا. ۱۳۹۹. کاربرد برنامه‌ریزی بیان ژن در پیش‌بینی خشک‌سالی (مطالعه موردی: ایستگاه سینوپتیک تبریز). مجله محیط‌زیست و مهندسی آب. ۵ (۱):۱-۱۴.

علی دادی ده کهنه، ص.، سلگی، ا.، شهنی دارابی، م. و زارعی، ح. ۱۳۹۸. ارزیابی مدل‌های ژنتیکی جهت مدل‌سازی جریان رودخانه. مهندسی آبیاری و آب ایران, ۹(۳۵): ۱-۱۷.

قبادیان، ر.، قربانی، م.ع. و خلج م. ۱۳۹۲. بررسی عملکرد روش برنامه‌ریزی بیان ژن درروند یابی سیلاب رودخانه زنگمار در مقایسه با روش موج دینامیکی. نشریه آب‌وخاک (علوم و صنایع کشاورزی) ۲۷ (۳): ۵۹۲-۶۰۲ .

کاوه کار، ش.، نعمتی، س. و ازانی. ع. ۱۳۹۲. ساختار درختی برنامه‌ریزی بیان ژن جهت شبیه‌سازی نوسانات تراز آب، کنفرانس بین‌المللی عمران، معماری و توسعه پایدار شهری، تبریز، دانشگاه آزاد اسلامی واحد تبریز.

مهدی زاده، س.، ۱۳۹۶. مدل‌سازی و پیش‌بینی برخی از پارامترهای مورد استفاده در مدیریت آب و کشاورزی با استفاده از مدل‌های هوش مصنوعی و تلفیق مدل‌های مذکور با روش‌های استوکاستیک. رساله دکتری تخصصی. دانشگاه ارومیه.

یونسی، م. و نوذری، ح. ۱۳۹۹. ارزیابی مدل‌های تلفیقی شبکه‌ی عصبی مصنوعی-موجک و برنامه‌ریزی بیان‌ژن-موجک در پیش‌بینی کردن خشک‌سالی کوتاه‌مدت. نشریه پژوهش‌های آبخیزداری (پژوهش و سازندگی). ۳۳(۱): ۳۹-۵۵.

Adib A., Mahmoudian Kafshgar Kalaee M., Mahmoudian Shoushtari M., Khalili K. 2017. Using of gene expression programming and climatic data for forecasting flow discharge by considering trend, normality, and stationarity analysis. Arabian Journal of Geosciences 10(9): 208.

Aytek A., Kisi O. 2008. A genetic programming approach to suspended sediment modeling. Journal of Hydrology 351(3-4): 288-298.

Azamathulla H.Md., Zahiri A. 2012. Flow discharge prediction in compound channels using linear genetic programming. Journal of Hydrology.454-455(6): 203-207.

Journal of Hydrology. 547: 348-364.

Salas J. D., Delleur j.w, Yevjevich v. and Lane,w.l. 1996. Applied Time Series in Hydrology, Mc Graw Hill.

Shiri J., Kisi O. 2011. Comparison of genetic programming with neuro-fuzzy systems for predicting short-term water table depth fluctuations. Comput. Geosci. 37: 1692-1701.

Traore S., Guven A. 2012. Regional-specific numerical models of evapotranspiration using gene-expression programming interface in Sahel. Water Resources. Management. 26: 4367-4380.

Wang Sh., Lian J., Peng Y., Hu B., Chen H. 2019. Generalized reference evapotranspiration models with limited climatic data based on random forest and gene expression programming in Guangxi, China. Agricultural Water Management. 221: 220–230.

Ferreira C. 2001. Gene expression programming: a new adaptive algorithm for solving problems.complex Systems.13 (2):87-129.

Ferreira C. 2006. Gene expression programming: Mathematical Modeling by an Artifical Intelligence. 2nd edn. Springer-Verlag Germany.463 p.

Lopes H.S., Weinert W.R. 2004. EGIPSYS: An enhanced gene expression programming approach for symbolic regression problems. International Journal of Applied Mathematics and Computer Science, 14(3): 375-384.

Moeeni H., Bonakdari H., Fatemi S.E., Zaji A.H. 2017. Assessment of Stochastic Models and a Hybrid Artificial Neural Network-Genetic Algorithm Method in Forecasting Monthly Reservoir Inflow. INAE Letters. 2(1): 13-23.

Moeeni, H., Bonakdari H., Fatemi, S.E, 2017. Stochastic model stationarization by eliminating the periodic term and its effect on time series prediction.

Iranian Journal of Irrigation and Drainage
No. 3, Vol. 15, Aug.-Sep. 2021, p. 582-597

نشریه آبیاری و زهکشی ایران
شماره۳، جلد ۱۵، مرداد - شهریور ۱۴۰۰، ص. ۵۹۷-۵۸۲

Assessment Effects of Data Preprocessing and Modeling Parameters of Gene Expression Programming on Accuracy of Time Series Forecasting

M. Salehi[1*], S.E. Fatemi[2]
Recived: Jan.19, 2021 Accepted: Apr.13, 2021

Abstract

Hydrological time–series is a time-dependent hydrological variable that finding the model of changes and predicting is the most important goal of time-series analysis. The purpose of this study is to simultaneously study the characteristics of time series and their prediction and the important parameters of the GEP for high-precision predictions in the training and validation. In this study, groundwater depth time-series of Chamchamal plain station located in Kermanshah province with a 12-year period and mountainous climate and the monthly time-series of Alaska temperature with a 50-year period and cold and dry climate have been used. Genexprotools5.0 software has been used to model time-series by GEP.

The results of studying with GEP showed that the periodicity of data properties that existed in the time series of temperature caused correlation results above 90% in different stages of training and validation. So that the effect of different parameters of GEP is less than 10% in improving results. On the other hand, by examining the time-series of groundwater depth, which lacks periodicity and has a descending ACF shape, the prediction results of the GEP with any effective expression parameter, R more than 44% in the validation wasn't obtained. This means that the time-series preprocessing has a greater impact on the prediction results. So that by eliminating the semester, the prediction results in all stages of modeling are significantly reduced. In this case, the best R for the validation is 50%.

Keywords: Forecasting, Gene Expression, Periodicity, Preprocessing, Time Series

1- Master of Water Resources, Water Engineering Department, Campus of Agricultur and Natural Resources, Razi University, Iran
2- Assistant Professor, Water Engineering Department, Campus of Agricultur and Natural Resources, Razi University, Iran
(*- Corresponding Author Email: e_fatemi78@yahoo.com)