نشریه حفاظت منابع آب و خاک، سال هشتم، شماره چهارم، تابستان ۱۳۹۸

ISSN 2251-7480

ارزیابی مدل برنامه‌ریزی بیان ژن برای برآورد بار رسوب معلق بر اساس پیش‌پردازش داده‌ها با روش آزمون گاما (مطالعه موردی: حوزه آبخیز رود زرد)

عادله علی جانپور شلمانی [۱*]، علی‌رضا واعظی [۲] و سید محمودرضا طباطبایی [۳]

۱) دانشجوی دکتری، گروه علوم خاک، دانشگاه زنجان، زنجان، ایران.

* نویسنده مسئول مکاتبات: adele.alijanpour@gmail.com

۲) استاد، گروه علوم خاک، دانشگاه زنجان، زنجان، ایران.

۳) استادیار پژوهشی، پژوهشکده حفاظت خاک و آبخیزداری، سازمان تحقیقات، آموزش و ترویج کشاورزی، تهران، ایران

تاریخ دریافت: ۱۳۹۷/۰۸/۲۰ تاریخ پذیرش: ۱۳۹۸/۰۱/۲۷

چکیده

یکی از عوامل مؤثر در شناخت مشکلات حوزه‌های آبخیز، بار رسوب معلق روزانه است. به‌دلیل نبود آمار کافی در اندازه‌گیری مستقیم بار رسوب معلق روزانه می‌توان از مدل‌های هوشمند مانند مدل برنامه‌ریزی بیان ژن (GEP) برای برآورد آن استفاده کرد. در این پژوهش، داده‌های ایستگاه هیدرومتری ماشین در حوزه آبخیز رود زرد، با طول دوره آماری ۳۶ سال (۱۳۵۶-۱۳۹۱) مورد استفاده قرار گرفت. متغیرهای ورودی به مدل GEP شامل دبی لحظه‌ای (Q)، متوسط دبی روزانه (Q_t) و متوسط بارندگی روزانه (P_t) به همراه سه گام تأخیر زمانی و متغیر خروجی به مدل شامل بار رسوب معلق روزانه می‌باشد. برای کاهش در وقت و هزینه، پیش‌پردازش داده‌های ورودی به مدل GEP با استفاده از روش آزمون گاما به‌دست آمد و به‌همراه ترکیبات بدون پیش‌پردازش (آزمون و خطا) وارد مدل GEP شد. نتایج مقایسه بین تمامی مدل‌ها نشان داد که برترین ترکیب متغیر ورودی حاصل از آزمون گاما، با کم‌ترین مقدار آماره‌ی خطای استاندارد برابر صفر، آماره‌ی گاما برابر ۰/۰۰۰۰۹۲ و آماره‌ی V_{ratio} برابر ۰/۰۱۲ و با ترکیب متغیرهای متوسط دبی روزانه به همراه دو گام تأخیر زمانی و متوسط بارندگی روزانه به همراه سه گام تأخیر زمانی، دقیق‌ترین و صحیح‌ترین برآورد را برای بار رسوب معلق داشت. این مدل دارای کم‌ترین مقدار RMSE=۱۶۷۱/۹۰ (ton/day) و MAE=۴۷۵/۶۸ (ton/day) و بیش‌ترین مقدار R^2=۰/۹۹ و NSE=۰/۹۹ در مقایسه با سایر مدل‌ها بود. بنابراین، استفاده از روش آزمون گاما به‌عنوان یک روش پیش‌پردازش داده‌ها توانست با انتخاب ترکیباتی از متغیرهای ورودی مناسب به مدل‌ها، به‌طور میانگین تا ۴۰ درصد مقدار خطای برآورد (RMSE) بار رسوب معلق روزانه را در مقایسه با ترکیبات ورودی حاصل از آزمون و خطا کاهش دهد و با افزایش تشابه بین مقادیر داده‌های مشاهداتی با داده‌های محاسباتی، عملکرد مدل GEP در برآورد بار رسوب معلق را افزایش دهد.

کلید واژه‌ها: حوزه‌ی آبخیز؛ رسوب؛ برنامه‌ریزی بیان ژن؛ آزمون گاما

مقدمه

یکی از شاخص‌های مهم در شناخت مشکلات حوزه‌های آبخیز، بار رسوب معلق روزانه (SSL[1]) می‌باشد. مسائلی مانند شدت فرسایش در نقاط بالادست حوضه و مقدار رسوبات حاصل در نقاط پایین دست و در پشت سدهای خاکی و بتنی، مشکلات مربوط به کیفیت آب رود و مسائل زیست محیطی، با دانستن مقدار دقیق SSL قابل توجیه است (Ouillon, 2018). برای اندازه‌گیری SSL می‌توان از روش‌های مستقیم و غیرمستقیم استفاده کرد. به دلیل مشکلات کمی در اندازه‌گیری مستقیم SSL مانند تعداد کم اندازه‌گیری‌ها به

[1] Suspended sediment load

دلیل نبود تعداد کافی دستگاه‌های اندازه‌گیری یا خرابی آن‌ها، نبود نیروی متخصص کافی و هزینه زیاد اندازه‌گیری و مشکلات کیفی که مربوط به زمان نمونه‌برداری در شرایطی غیر از شرایط سیلابی می‌باشد، از روش‌های غیرمستقیم برای برآورد آن استفاده می‌شود (Melesse *et al.*, 2011 ،Kisi and Ozkan, 2017). از بین روش‌های غیرمستقیم، روش‌های مبتنی بر هوش مصنوعی قادر هستند تا پاسخ و رفتار مناسبی از حوزه‌های آبخیز در برابر متغیرهای ورودی را ، برای برآورد SSL ارائه دهند (Demirci and Baltaci, 2013). در این تحقیق از بین روش‌های مبتنی بر هوش مصنوعی برای برآورد SSL، مدل برنامه‌ریزی بیان ژن (GEP[1]) استفاده شد. مدل GEP به‌عنوان مدلی بر پایه‌ی الگوریتم تکاملی، برای برآورد پدیده‌های غیرخطی مناسب می‌باشد (Anganibi *et al.*, 2014، Muzzammil *et al.*, 2015).

محققان در مطالعات خود برای برآورد SSL حوزه‌های آبخیز، از روش‌های هوش مصنوعی و سایر روش‌ها استفاده کردند. به‌عنوان مثال، در مطالعات انجام شده توسط Guven و Talu (2010)، در حوزه‌ی آبخیز فرات میانی در ترکیه، مقدار SSL با استفاده از مدل‌های GEP، رگرسیون خطی و منحنی سنجه رسوب برآورد شد. در این تحقیق از داده‌های ایستگاه هیدرومتری هنیس کریک مربوط به سال‌های ۲۰۰۵ تا ۲۰۰۷ که شامل ۶۶ درصد از داده‌ها بود، برای آموزش مدل‌ها و از داده‌های مربوط به سال‌های ۲۰۰۷ تا ۲۰۰۸ (۳۳ درصد کل داده‌ها) برای آزمون مدل‌ها استفاده شد. در این تحقیق متغیرها برای ورود به مدل‌ها در دو گروه داده‌های دبی روزانه و رسوب متناظر با آن و همچنین دبی روزانه و رسوب با تأخیر سه روز، قرار گرفتند. انتخاب این ترکیبات به‌صورت دستی و تصادفی انجام شد. نتایج این تحقیق نشان داد که مدل GEP با R^2= ۰/۹۹ و (gr/l) RMSE=۳۴۸/۵۰ کارآمدترین مدل در برآورد SSL در مقایسه با سایر مدل‌ها بود. barzegari و همکاران (2015)، برای برآورد بار رسوب معلق در حوزه‌های آبخیز لرستان از روش‌های شبکه عصبی مصنوعی، درخت تجزیه و منحنی سنجه رسوب استفاده کردند. متغیرهای مورد استفاده در این تحقیق در بازه زمانی ۱۹۹۶ تا ۲۰۰۶، دبی روزانه و رسوب متناظر با آن بود. نتایج نشان داد که مدل هوشمند شبکه عصبی مصنوعی در مقایسه با سایر مدل‌ها و با NSE=۱ و RMSE=۰/۰۰۳۱ توانست مقدار SSL روزانه را در ایستگاه آزنا را با صحت بالاتری برآورد کند. Emamgholizadeh و Karimi Demneh (2018)، برای برآورد SSL در حوزه‌ی آبخیز تالار از روش‌های شبکه عصبی مصنوعی، GEP و منحنی سنجه‌ی رسوب استفاده کردند. داده‌های مورد استفاده در این تحقیق، دبی با زمان تأخیر دو روز و رسوب معادل آن بود. داده‌ها به‌صورت روش آزمون و خطا وارد مدل‌ها شدند. نتایج نشان داد که مدل GEP با R^2=۰/۷۵ و (ton/day) ۱۲۶۹/۷ =MAEتوانست در مقایسه با مدل شبکه عصبی مصنوعی با R^2=۰/۶۱ و (ton/day) MAE=۳۰۲۳/۴۵ و مدل منحنی سنجه رسوب با R^2=۰/۳۹ و (ton/day) MAE=۶۷۳۲/۸، برآورد دقیق‌تری از SSL را داشته باشد.

همان‌طور که مشاهده شد در این تحقیقات و بسیاری از مطالعات انجام شده در زمینه‌ی برآورد بار رسوب معلق روزانه، متغیرهای ورودی به‌صورت روش آزمون و خطا وارد مدل‌ها شده‌اند و نقش انتخاب ترکیب متغیرهای ورودی بهینه به مدل‌ها در برآورد دقیق‌تر SSL در نظر گرفته نشده است و همچنین، در بیشتر مطالعات نقش متغیر بارندگی به عنوان عامل مؤثر در برآورد بار رسوب معلق در نظر گرفته نشده و تنها از متغیر دبی برای برآورد آن استفاده شده است. در این تحقیق، در مدل‌سازی بار رسوب معلق روزانه با استفاده از مدل GEP، تعیین مهم‌ترین و تأثیرگذارترین داده‌ها در برآورد SSL، از تصمیمات مهم در فرآیند توسعه این مدل می‌باشد. در طی فرآیند مدل‌سازی، استفاده از تمام متغیرهای ممکن و

[1] Genetic Expression Programing

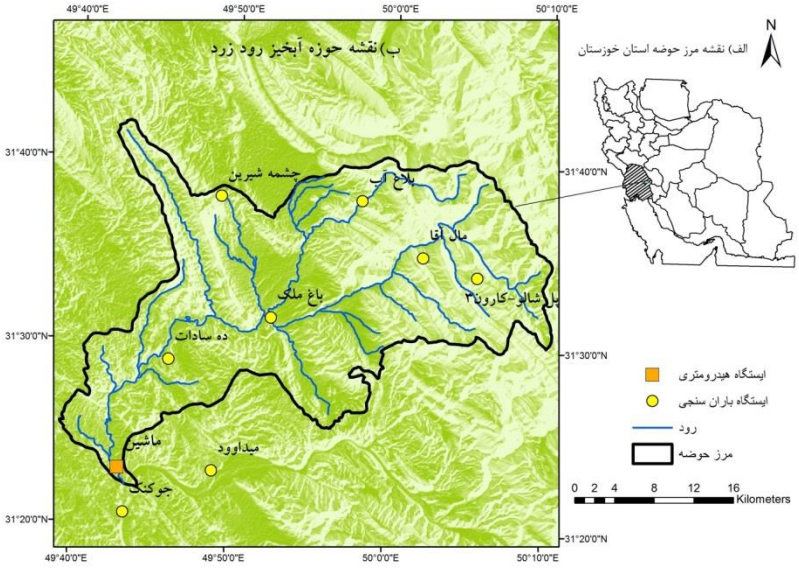
موجود، می‌تواند پیچیدگی مدل را افزایش دهد در صورتی‌که، استفاده از متغیرهای ورودی درست و منطقی که شامل اطلاعات کافی باشد، برآورد دقیق‌تری از متغیر مورد نظر را خواهد داشت (Galelli *et al.*, 2014). زمانی‌که متغیرها به‌صورت روش آزمون و خطا وارد مدل-ها شوند، ممکن است متغیری که تأثیر کمی در برآورد درست متغیر خروجی دارد در فرآیند مدل‌سازی به‌کار رود و متغیرهای تأثیرگذار حذف گردند. این نوع مدل-سازی توجیه و توضیح نتایج مدل‌ها را با مشکل مواجه می‌کند (Wu *et al.*, 2014). یکی از روش‌های که برای به‌دست آوردن ترکیبات ورودی بهینه به مدل‌ها مورد استفاده قرار می‌گیرد، روش آزمون گاما می‌باشد. این روش می‌تواند با انتخاب متغیرهای مناسب و تولید مدل کارآمد، پاسخ منطقی و صحیح از مدل در برآورد SSL را ارائه دهد (Jajarmizadeh *et al.*, 2015). اهداف این تحقیق شامل موارد ذیل می‌باشد:

- استفاده از روش آزمون گاما برای به‌دست آوردن بهترین و بهینه‌ترین ترکیبات متغیری برای ورود به مدل GEP،
- ارزیابی دقت و صحت مدل GEP در برآورد SSL، با مقایسه‌ی نتایج مدل‌سازی با ترکیبات متغیر ورودی به مدل‌ها با استفاده از روش پیش‌پردازش آزمون گاما و روش بدون پیش‌پردازش آزمون و خطا،
- استفاده از متغیر دینامیک بارندگی به‌همراه دبی برای بررسی تأثیر آن در میزان دقت برآورد SSL حوزه‌ی آبخیز رود زرد.

مواد و روش‌ها

منطقه مورد مطالعه

حوزه آبخیز رود زرد با دامنه‌ی تغییرات ارتفاع حدود ۴۰۰ تا ۳۳۰۰ متر از سطح دریا، در جنوب غربی ایران و شرق استان خوزستان بین طول‌های شرقی ´ ۳۹ °۴۹ تا ´ ۱۰ °۵۰ و عرض‌های شمالی ´ ۲۱ °۳۱ تا ´ ۴۱ °۳۱ واقع شده است (شکل ۱). این حوضه یکی از زیرحوضه‌های رودخانه‌ی مارون-جراحی است که از ارتفاعات جنوب غرب زاگرس به‌وسیله‌ی رودخانه‌های الله و مارون زهکشی می‌شود. مساحت این حوضه ۸۶۱ کیلومتر مربع، محیط آن ۱۷۳ کیلومتر، ارتفاع متوسط از سطح دریا ۱۵۲۴ متر، طول آبراهه اصلی ۴۷/۵ کیلومتر، شیب متوسط رودخانه ۲۷/۴ درصد، ضریب گراویلیوس ۱/۷، ضریب شکل ۲/۶ و نسبت انشعابات آبراهه‌ای ۱/۶ می‌باشد. اقلیم غالب منطقه براساس روش دومارتن، نیمه‌خشک است.



شکل ۱. نقشه حوزه‌ی آبخیز رود زرد

داده‌های مورد استفاده

در این پژوهش، ایستگاه هیدرومتری ماشین (از نوع درجه‌ی یک) در حوزه آبخیز رود زرد، مورد استفاده قرار گرفت. دوره‌ی داده‌های آماری قابل استفاده از ایستگاه هیدرومتری ماشین با توجه به داده‌های دینامیک حوزه آبخیز، ۳۶ سال (۱۳۵۶-۱۳۹۱) است که ۸۰ درصد داده‌ها برای آموزش و ۲۰ درصد داده‌ها برای آزمون مدل‌ها مورد استفاده قرار گرفت. داده‌های مورد استفاده در این تحقیق، از سازمان تماب اخذ شد. این داده‌ها شامل متغیرهای ورودی به مدل GEP شامل دبی لحظه‌ای (Q)، متوسط دبی روزانه (Q_i) و متوسط بارندگی روزانه (P_i) و متغیر خروجی به مدل شامل بار رسوب معلق روزانه (SSL) می‌باشد. برای افزایش دقت و کارآیی مدل GEP در برآورد SSL، متغیرهای Q_i و P_i با زمان تأخیر تا سه روز قبل مورد استفاده قرار گرفتند. متوسط بارندگی روزانه در طول دوره آماری در این حوضه ۲/۳۵ میلی‌متر و متوسط دبی روزانه ۱۰/۲۱ (m^3/s)، متوسط دبی لحظه‌ای ۹/۳۹ (m^3/s) و متوسط بار رسوب معلق ۱۲۱۸/۴۳ (ton/day) است. مشخصات آماری داده‌های مورد استفاده در این تحقیق در جدول ۱ آورده شده است.

جدول ۱. مشخصات آماری داده‌های مورد استفاده در حوزه‌ی آبخیز رود زرد

نوع داده	کمینه	بیشینه	میانگین	ضریب تغییرات
دبی لحظه‌ای Q (m^3/s)	۰/۲۲	۱۰۹	۹/۳۹	۱/۵۰
دبی متوسط روزانه Q_i (m^3/s)	۰/۲۰	۱۷۴	۱۰/۲۱	۱/۸۹
دبی متوسط روزانه یک روز قبل Q_{i-1} (m^3/s)	۰/۲۰	۳۷۰	۱۱/۷۳	۲/۵۵
دبی متوسط روزانه دو روز قبل Q_{i-2} (m^3/s)	۰/۱۵	۵۳۳	۹/۹۸	۲/۸۴
دبی متوسط روزانه سه روز قبل Q_{i-3} (m^3/s)	۰/۱۵	۴۴۳	۹/۷۴	۲/۶۰
بارندگی متوسط روزانه P_i (mm)	۰/۰۰	۵۵/۵۰	۲/۳۵	۳/۳۱
بارندگی متوسط روزانه یک روز قبل P_{i-1} (mm)	۰/۰۰	۹۱/۲۳	۲/۸۵	۳/۴۵
بارندگی متوسط روزانه دو روز قبل P_{i-2} (mm)	۰/۰۰	۷۲/۸۵	۱/۹۲	۳/۴۶
بارندگی متوسط روزانه سه روز قبل P_{i-3} (mm)	۰/۰۰	۵۱/۲۹	۱/۵۷	۳/۵۹
بار رسوب معلق روزانه SSL ($ton\ day^{-1}$)	۰/۳۸	۵۷۳۰۰/۴۶	۱۲۱۸/۴۳	۴/۸۶

آزمون گاما (GT)[1]

در فرآیند مدل‌سازی انتخاب ترکیب متغیرهای ورودی به مدل‌ها با استفاده از روش آزمون و خطا، مشکل، وقت‌گیر و خسته‌کننده است. اگر تعداد متغیرهای ورودی به مدل برابر با m متغیر باشد، تعداد 2^m-1 حالت برای ترکیبات متغیر ورودی به مدل وجود دارد. بنابراین بررسی تمامی این ترکیبات، زمانی که تعداد متغیرهای ورودی زیاد باشد، غیر ممکن به‌نظر می‌رسد. به‌منظور رفع این مشکل می‌توان از روش آزمون گاما به‌عنوان یک روش پیش‌پردازش داده‌ها قبل از ورود به مدل، استفاده کرد. آزمون گاما به‌عنوان یک تکنیک شناسایی متغیرهای ورودی تأثیرگذار در برآورد متغیر خروجی است که نخستین بار توسط Stefansson و همکاران (1997) کشف شد و بعدها توسط سایر محققان (Malik *et al*, 2017, Jamalizadeh *et al*., 2008, Shamim *et al*., 2016) مورد استفاده قرار گرفت. در این روش، برای تمام ترکیبات تعریف شده با توجه به فرمول 2^m-1، مقدار حداقل خطای استاندارد، آماره‌ی گاما و آماره‌ی V_{ratio} محاسبه می‌شود. به‌منظور محاسبه‌ی آماره‌ی گاما (Γ)، خط رگرسیون به‌صورت زیر تعریف می‌شود (Malik *et al*, 2017):

$$Y = A\delta + \Gamma \qquad (1)$$

که در آن؛ Y بردار خروجی خط رگرسیون، A شیب خط رگرسیون و δ متغیر مستقل است. مقدار آماره‌ی گاما در واقع عرض از مبدا خط رگرسیون است. مقدار شیب خط،

[1] Gamma Test

آن مقدار از واریانس خروجی را که مدل قادر به برآورد آن نیست نشان می‌دهد. هرچقدر مقدار آماره‌ی گاما به صفر نزدیک‌تر باشد، ترکیب متغیری حاصل، برآورد دقیق‌تری از متغیر خروجی را خواهد داشت و محدودیتی برای ساخت مدل مناسب وجود ندارد (Shamim *et al.*, 2016). آماره‌ی V_{ratio}، خروجی مطلوب را برای توابع مدل‌سازی می‌کند (رابطه ۲).

$$V_{ratio} = \frac{\Gamma}{\sigma^2(y)} \qquad (۲)$$

که در آن: $\sigma^2(y)$ واریانس خروجی از y است. این نسبت بررسی شکلی را که وابسته به دامنه خروجی است امکان-پذیر ساخته و قادر است تا خروجی مناسب را برای توابع هموار مدل‌سازی کند. در آزمون گاما، مقدار آماره V_{ratio} بین صفر و یک تغییر می‌کند. هرچقدر مقدار این آماره به یک نزدیک‌تر باشد، نشان می‌دهد که متغیرهای ورودی انتخاب شده قادر به ارائه خروجی مطلوب نمی‌باشند. درصورتی‌که مقدار صفر یا نزدیک به صفر آماره V_{ratio}، نشان‌دهنده‌ی قدرت متغیرهای ورودی انتخابی به‌عنوان متغیرهای بهینه در برآورد دقیق متغیر خروجی می‌باشد (Jajarmizadeh *et al.*, 2015). در این مطالعه، برای به-دست آوردن بهترین ترکیب‌های متغیرهای ورودی به مدل GEP، از بسته‌ی نرم‌افزاری آزمون گاما موجود در نرم‌افزار WinGamma™ استفاده شد.

برنامه‌ریزی بیان ژن

مدل برنامه‌ریزی بیان ژن (GEP)، یک روش بهینه‌سازی فرا ابتکاری برای رسیدن به برآورد درست از متغیر هدف است. این مدل با الهام گرفتن از طبیعت به سمت تکامل و بهینه مطلق پیش می‌رود (Bagatur and Onen, 2014). این روش براساس الگوریتم‌های گردشی و برمبنای نظریه تکاملی داروین است که توسط فریرا در سال ۲۰۰۱ ارائه شد (Ferreira, 2001). روش GEP ترکیبی از الگوریتم ژنتیک (GA[1]) و برنامه‌نویسی ژنتیک (GP[2]) است که ژنوتایپ کروموزوم‌ها در مدل GEP مشابه با GA دارای یک ساختار خطی با طول ثابت و فنوتایپ کروموزوم‌ها به‌صورت یک ساختار درختی با طول و اندازه‌ی متفاوت مشابه با GP است (Ferreira, 2001). به‌دلیل اینکه در مدل GEP تمام ساختارهای درختی با اندازه و اشکال متفاوت، در کروموزوم‌های خطی با طول ثابت کدگذاری می‌شوند، این امر سبب جداسازی کامل فنوتایپ و ژنوتایپ شده، در نتیجه باعث می‌گردد تا سیستم بتواند از تمام مزایای تکاملی بهره‌مند شود (Guven and Talu, 2010). در این روش ابتدا بهینه‌سازی شامل فرآیندهای جهش، وارونگی، تولید مثل و انتخاب بهترین ژن، در ساختار خطی انجام شده و سپس به‌صورت ساختار درختی بیان می‌شود. این مسئله باعث می‌شود تا تنها ژنوم‌های اصلاح شده به نسل بعد منتقل شوند. بنابراین، نیازی به ساختارهای سنگین برای تکثیر و جهش وجود نخواهد داشت (Ferreira, 2001). در این روش متغیرهای هدف با استفاده از مجموعه‌ای از توابع و ترمینال‌ها، مدل‌سازی می‌شوند. مجموعه توابع معمولا شامل توابع اصلی (+, -, ×, /)، توابع مثلثاتی (x2, exp, log) یا توابع تعریف شده توسط کاربر است که احتمال می‌دهد برای تفسیر مدل مناسب خواهد بود (Azamathulla, 2013). مجموعه‌ی ترمینال شامل متغیرهای مستقل و ثابت‌های عددی است. در این تحقیق، به منظور مدل‌سازی با استفاده از روش GEP از نرم‌افزار GEPXpro Tools5.0 استفاده شد.

شاخص‌های آماری سنجش مدل‌ها

در این تحقیق، عملکرد مدل‌های برنامه‌ریزی بیان ژن با استفاده از ترکیبات متغیرهای ورودی با روش آزمون گاما در مقایسه با روش آزمون و خطا برای برآورد بار رسوب معلق روزانه، با استفاده از شاخص‌های آماری ضریب تبیین (R^2[3])، ریشه میانگین مربعات خطا (RMSE[4])، میانگین قدر مطلق خطا (MAE) و معیار نش- ساتکلیف (NSE) مورد ارزیابی قرار گرفت:

[1] Genetic algorithms
2 Genetic programming

3 Coefficient of Determination
[4] Root Mean Square error

$$R^2 = \left[\frac{\sum_{i=1}^{n}(s_O-\overline{s_O})(s_p-\overline{s_p})}{\sqrt{\sum_{i=1}^{n}(s_O-\overline{s_O})^2\sum_{i=1}^{n}(s_p-\overline{s_p})^2}}\right]^2 \quad (۳)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(s_p - s_O)^2} \quad (۴)$$

$$MAE = \frac{\sum_{i=1}^{n}|(s_O-s_p)|}{n} \quad (۵)$$

$$NSE = 1 - \frac{\sum_{i=1}^{n}(s_p-s_O)^2}{\sum_{i=1}^{n}(s_O-\overline{s_O})^2} \quad (۶)$$

که در آن‌ها: s_O و s_p به‌ترتیب بار رسوب معلق مشاهده‌ای و برآورد شده، $\overline{s_O}$ میانگین بار رسوب معلق مشاهده‌ای، $\overline{s_p}$ میانگین بار رسوب معلق برآورد شده و n تعداد داده‌ها می‌باشد

نتایج و بحث

نتایج آزمون گاما

در روش آزمون گاما با توجه به نه متغیر ورودی تعریف شده ($Q,Q_i,Q_{i-1},Q_{i-2},Q_{i-3},P_i,P_{i-1},P_{i-2},P_{i-3}$) در این تحقیق، تعداد ۵۱۱ ترکیب مورد بررسی قرار گرفت و پنج ترکیب اول این آزمون به‌عنوان بهینه‌ترین ترکیبات متغیری با کمترین مقدار خطای استاندارد، آماره‌ی گاما و آماره‌ی V_{ratio}، برای مقایسه با ترکیبات حاصل از روش آزمون و خطا (روش بدون پیش‌پردازش داده‌ها)، به مدل GEP معرفی شدند. این مقایسه به‌دلیل بررسی صحت تأثیر ترکیبات حاصل از آزمون گاما به عنوان یک روش پیش-پردازش داده‌ها در عملکرد مدل GEP انجام شد. جدول ۲ مقدار آماره‌های آزمون گاما (GT) برای بهترین ترکیبات متغیر ورودی به مدل GEP را نشان می‌دهد. نتایج این جدول نشان داد که بهترین ترکیب متغیر ورودی برای حوزه آبخیز رود زرد، با حداقل مقدار آماره خطای استاندارد برابر صفر، آماره‌ی گاما برابر ۰/۰۰۰۰۹۲ و آماره‌ی V_{ratio} برابر ۰/۰۱۲، ترکیب با متغیرهای دبی متوسط روزانه (Q_i)، دبی متوسط روزانه یک روز قبل (Q_{i-1})، دبی متوسط روزانه دو روز قبل (Q_{i-2})، بارندگی متوسط روزانه (P_i)، بارندگی متوسط روزانه یک روز قبل (P_{i-1})، بارندگی متوسط روزانه دو روز قبل (P_{i-2}) و بارندگی متوسط روزانه سه روز قبل (P_{i-3}) بود. در تحقیقات مختلف نشان داده شده است که استفاده از ترکیبات متغیر ورودی به مدل‌ها برای برآورد متغیرهای خروجی با استفاده از روش GT، نتایج دقیق‌تر و قابل قبولی را ارائه می‌دهد (Noori *et al*., 2009, Remesan *et al*., 2008, Wan Jaafar *et al*., 2011). در تحقیقی مشابه با این تحقیق در برآورد بار رسوب معلق، Rashidi و همکاران (2016) از مدل هوشمند شبکه عصبی مصنوعی در حوزه‌ی آبخیز کورکورسر در شمال ایران استفاده کردند. این محققان برای به‌دست آوردن بهترین ترکیب متغیر برای ورود به مدل از روش GT استفاده کردند. نتایج نشان داد که استفاده از روش آزمون گاما باعث افزایش دقت مدل در برآورد بار رسوب معلق با R^2=۰/۸۸، RMSE=۱۴/۰۴۵ (ton/day) و NSE=۰/۸۸ در مقایسه با ترکیبات ورودی بدون پیش‌پردازش با R^2=۰/۷۹، RMSE=۱۸/۳۶ (ton/day) و NSE=۰/۷۳ شد.

جدول ۲. نتایج بهترین ترکیبات متغیر ورودی به مدل GEP با استفاده از آزمون گاما

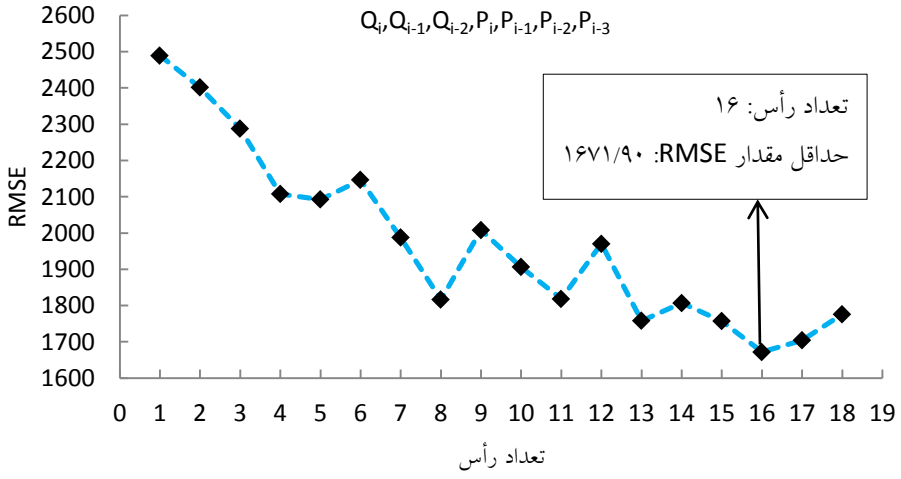
شماره مدل	ترکیب‌های ورودی	V_{Ratio}	گاما	خطای استاندارد
۱	$Q_i,Q_{i-1},Q_{i-2},P_i,P_{i-1},P_{i-2},P_{i-3}$	۰/۰۱۲	۰/۰۰۰۰۹۲	۰/۰۰
۲	$Q,Q_i,Q_{i-1},P_i,P_{i-1},P_{i-2},P_{i-3}$	۰/۰۲۲	۰/۰۰۰۱	۰/۰۰
۳	$Q,Q_i,P_i,P_{i-1},P_{i-2},P_{i-3}$	۰/۰۲۳	۰/۰۰۰۲	۰/۰۰۰۳
٤	$Q,Q_i,Q_{i-2},P_i,P_{i-1},P_{i-2},P_{i-3}$	۰/۰٤۷	۰/۰۰۰۳	۰/۰۰۰۲
٥	$Q,Q_i,Q_{i-1},Q_{i-2},P_i,P_{i-1},P_{i-2},P_{i-3}$	۰/۱۱٤	۰/۰۰۰۸	۰/۰۰۰۳

نتایج مدل برنامه‌ریزی بیان ژن

جدول ۳ مقادیر پارامترهای مورد استفاده در روش GEP، برای برآورد SSL در حوزه‌ی آبخیز رود زرد را نشان می‌دهد. معیار تابع خطای برازش برای برآورد SSL در این حوضه، RMSE در نظر گرفته شد. تعداد رأس برای مدل‌های GEP با ترکیبات متغیری مختلف در دو گروه داده‌ای حاصل از روش آزمون و خطا و همچنین روش آزمون گاما، به کمک سعی و خطا، برابر با ۱۰ تا ۱۸ تعیین شد. شکل ۲ واسنجی تعداد رأس با روش سعی و خطا برای ترکیب متغیر ورودی $Q_i,Q_{i-1},Q_{i-2},P_i,P_{i-1},P_{i-2},P_{i-3}$ حاصل از روش آزمون گاما را نشان می‌دهد. با توجه به شکل ۲ مشاهده شد که تعداد رأس مناسب برای رسیدن به حداقل مقدار خطا (ton/day) RMSE=۱۶۷۱/۹۰ در این ترکیب، برابر با ۱۶ رأس است. همچنین در این تحقیق در مدل‌سازی بار رسوب معلق روزانه با استفاده از روش GEP مشاهده شد که ارتباط منطقی بین متغیرهای ورودی با متغیر خروجی به‌صورت تابع پیوند + است که Emamgholizadeh و همکاران (2015) نیز در تحقیقات خود به نتیجه مشابه دست یافتند. این تابع به‌عنوان عملگر ریاضی بیان ژن در بالاترین سطر از ساختار درختی به‌عنوان تابع اتصال ژن‌ها انتخاب شد. استفاده از این تابع باعث افزایش سرعت و کیفیت الگوریتم در طی مراحل اجرای مدل شد.

جدول ۳. مقادیر پارامترهای مدل GEP در برآورد SSL

پارامتر	مقدار	پارامتر	مقدار
تعداد کروموزوم‌ها، Number of chromosomes	۳۰	نرخ ترکیب تک نقطه‌ای، One-point recombination rate	۰/۳
تعداد ژن‌ها، Number of Genes	۳	نرخ ترکیب دو نقطه‌ای، Two-point recombination rate	۰/۳
عملگر ریاضی بیان ژن، Linking function	+	نرخ ترکیب ژن، Gene recombination rate	۰/۱
نرخ جهش، Mutation rate	۰/۰۴۴	تعداد رأس، Number of head	۱۰-۱۸
نرخ وارونگی، Inversion rate	۰/۱	تعداد جمعیت تولیدی	۱۰۰۰
تعداد ژن‌ها، Number of Genes	۳	ترانهش ژنی، Gene transposition rate	۰/۱



شکل ۲. نمودار واسنجی تعداد رأس با روش سعی و خطا برای ترکیب $Q_i,Q_{i-1},Q_{i-2},P_i,P_{i-1},P_{i-2},P_{i-3}$ حاصل از روش GT

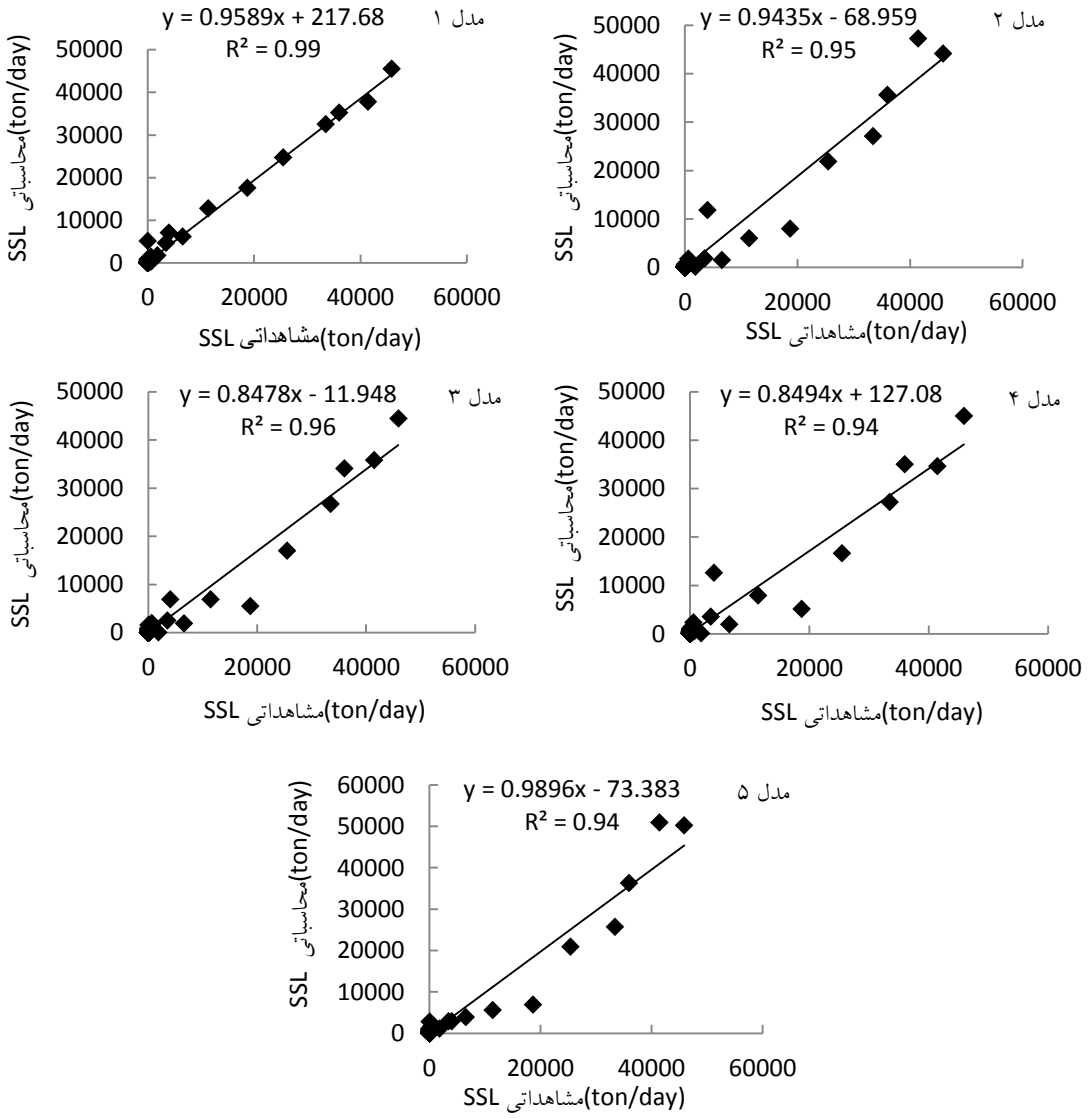
جدول ۴ نتایج برآورد SSL با استفاده از ترکیبات بهینه‌ی متغیر حاصل از آزمون گاما و ترکیبات متغیری به روش آزمون و خطا برای ورود به مدل GEP برای مجموعه داده‌های آزمون را نشان می‌دهد.

جدول ۴. نتایج مدل‌های GEP با استفاده از ترکیبات متغیری حاصل از آزمون گاما و روش آزمون و خطا در برآورد SSL

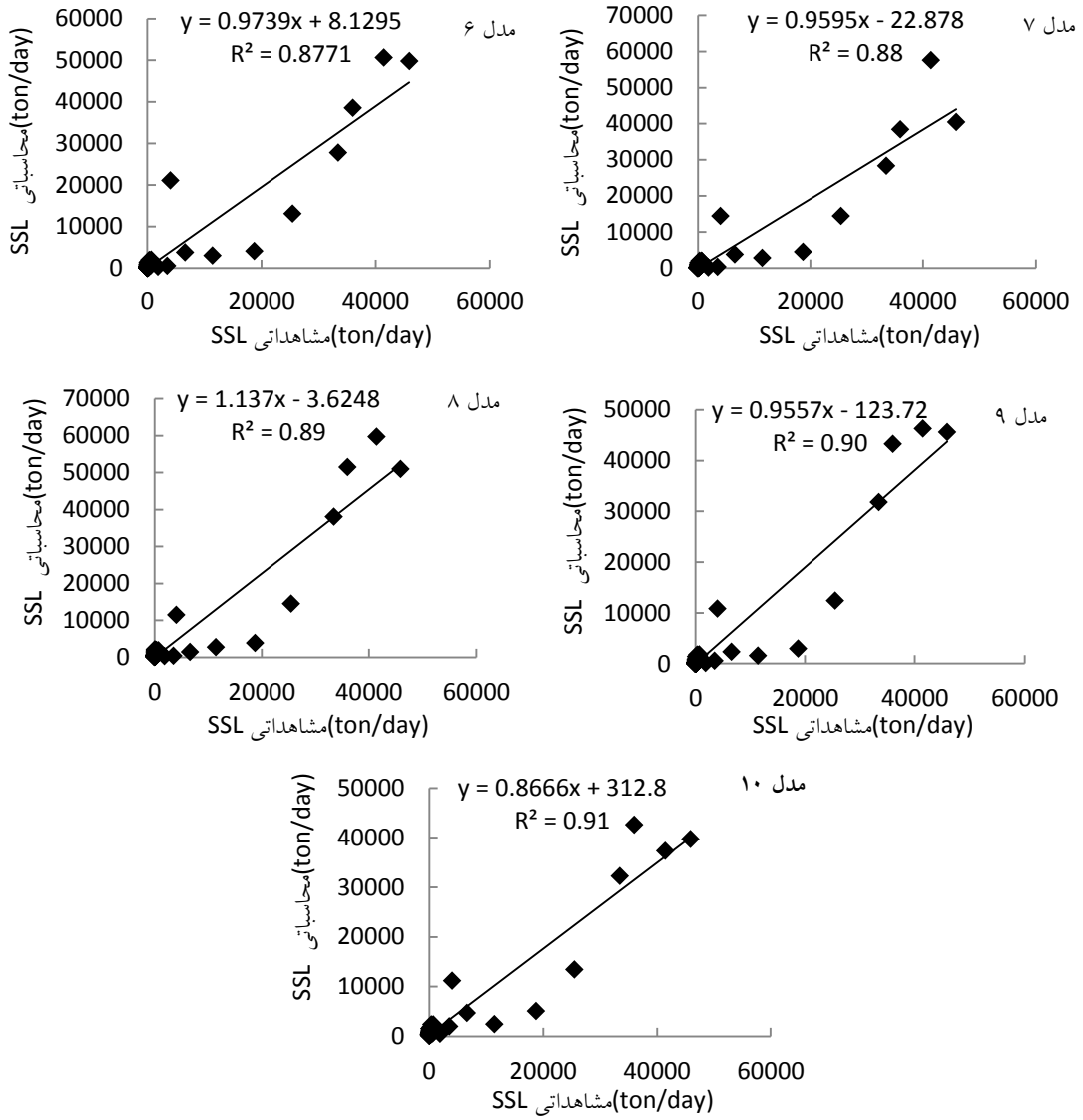
شماره مدل	ترکیب‌های ورودی	روش ورودی به مدل	MAE	RMSE	NSE	R^2
۱	$Q_i,Q_{i-1},Q_{i-2},P_i,P_{i-1},P_{i-2},P_{i-3}$	آزمون گاما	٤٧٥/٦٨	١٦٧١/٩٠	۰/۹۹	۰/۹۹
۲	$Q,Q_i,Q_{i-1},P_i,P_{i-1},P_{i-2},P_{i-3}$	آزمون گاما	٧٥٥/٧٧	٢٠٦٩/٩٨	۰/۹٥	۰/۹٥
۳	$Q,Q_i,P_i,P_{i-1},P_{i-2},P_{i-3}$	آزمون گاما	٨٥٢/٤٢	٢٢٨٧/٧٩	۰/۹٤	۰/۹٦
٤	$Q,Q_i,Q_{i-2},P_i,P_{i-1},P_{i-2},P_{i-3}$	آزمون گاما	٨٤١/٢٤	٢٢٠٦/٦١	۰/۹۳	۰/۹٤
٥	$Q,Q_i,Q_{i-1},Q_{i-2},P_i,P_{i-1},P_{i-2},P_{i-3}$	آزمون گاما	٩٥٧/٨٧	٢٤٣١/٦٥	۰/۹٤	۰/۹٤
٦	Q,Q_i	آزمون و خطا	١٢٤٢/٠١	٣٤٦٠/٣٩	۰/۸٧	۰/۸٧
٧	Q,Q_i,Q_{i-1},Q_{i-2}	آزمون و خطا	١٢٥٧/٤٠	٣٣٨٧/٦٩	۰/۸٧	۰/۸۸
۸	$Q,Q_i,Q_{i-1},Q_{i-2},Q_{i-3}$	آزمون و خطا	١٧٢٠/٢١	٤٢٢١/١٨	۰/۸۳	۰/۸۹
۹	$Q,Q_i,Q_{i-1},P_i,P_{i-1}$	آزمون و خطا	١٠٧٣/٢١	٣٠٠٨/٩٦	۰/۹۰	۰/۹۰
۱۰	$Q,Q_i,Q_{i-1},Q_{i-2},Q_{i-3},P_i,P_{i-1},P_{i-2}$	آزمون و خطا	١٢٧١/٦٢	٢٨٢٨/٨٠	۰/۹۱	۰/۹۱

نتایج جدول ۴ نشان داد، مدل ۱ با متغیرهای ورودی شامل متوسط دبی روزانه به همراه دو گام تأخیر زمانی و متوسط بارندگی روزانه به همراه سه گام تأخیر زمانی توانست مقدار SSL را با دقت بالا ((ton/day) RMSE=۱۶۷۱/۹۰ و NSE=۰/۹۹) برآورد کند. این مدل، بهینه‌ترین ترکیب حاصل از آزمون گاما با کمترین مقدار آماره‌ی خطای استاندارد، آماره‌ی گاما و آماره‌ی V_{ratio} نیز بود. همچنین، براساس نتایج جدول ۴ مربوط به ترکیب‌های متغیرهای ورودی حاصل از روش آزمون و خطا (مدل ۶ تا ۱۰)، ترکیب با متغیرهای دبی لحظه‌ای، متوسط دبی روزانه به همراه سه گام تأخیر زمانی و متوسط بارندگی روزانه به همراه دو گام تأخیر زمانی (مدل ۱۰)، بهترین مدل GEP برای برآورد SSL با استفاده از ترکیبات بدون پیش‌پردازش بود. مقدار RMSE این مدل برابر ۲۸۲۸/۸۰ (ton/day) و مقدار NSE آن برابر ۰/۹۱ می‌باشد. همچنین، در این جدول مشاهده شد مقدار خطای برآورد در استفاده از متغیر متوسط بارندگی روزانه به همراه متغیر متوسط دبی روزانه (مدل ۹ و ۱۰) در ترکیبات ورودی با روش آزمون و خطا در مقایسه با زمانی که متغیر متوسط دبی روزانه به تنهایی استفاده می‌شود (مدل ۶، ۷ و ۸)، به‌طور چشمگیری کاهش یافته است. مقایسه نتایج مدل‌ها با ترکیبات متغیرهای ورودی با استفاده از روش پیش‌پردازش آزمون گاما و روش بدون پیش‌پردازش آزمون و خطا در جدول ۴ نشان داد، تمامی ترکیبات متغیری بهینه حاصل از GT برای ورود به مدل GEP دارای مقادیر RMSE و MAE کم‌تر، R^2 و NSE بالاتر و میانگین ۴۰ درصد مقدار خطای برآورد بار رسوب معلق روزانه (RMSE) کمتر در مقایسه با ترکیبات ورودی حاصل از روش آزمون و خطا بود. این امر نشان دهنده‌ی کارآیی بالا و صحیح مدل‌های GEP در برآورد SSL با استفاده از این ترکیبات، در مقایسه با ترکیباتی است که به‌صورت آزمون و خطا وارد مدل شده‌اند. بنابراین در این تحقیق استفاده از روش آزمون گاما به‌عنوان یک روش پیش پردازش توانست متغیرهای مهم و تأثیرگذار در برآورد SSL را شناسایی کند و با ایجاد ترکیبات ورودی مناسب به مدل GEP، دقت برآورد را به میزان قابل توجهی افزایش دهد. در این روش مقدار میانگین مربعات خطای مدل پیش از استفاده محاسبه شده

و متغیرهای ورودی بی‌ربط یا در معرض خطای اندازه-گیری زیاد، حذف می‌شود. بنابراین، این روش به‌عنوان یک روش ریاضی نسبتا سریع قادر است تا منتخبی از ورودی‌ها که مقدار تقریبی آماره‌ی گامای آن‌ها حداقل است را پیدا کرده و در نهایت بهترین ترکیب از متغیرهای ورودی به مدل غیرخطی را رائه دهد. نتایج فوق در تطابق با نتایج Malik و همکاران (۲۰۱۷) می‌باشد. این محققان برای یافتن بهترین ترکیب متغیرهای ورودی به مدل‌های هوشمند از آزمون گاما استفاده کردند. نتایج نشان داد که استفاده از این ترکیبات مقدار R^2 مدل‌ها در برآورد غلظت معلق روزانه در رودخانه پارنهیتا در هند را از ۰/۹۰ در ترکیبات دستی به مقدار ۰/۹۹ با ترکیبات حاصل از GT افزایش داد. همچنین در تحقیقی دیگر Jajarmizadeh و همکاران (۲۰۱۵) نشان دادند که استفاده از روش GT برای ورود ترکیبات متغیری به مدل ماشین بردار پشتیان در مقایسه با ترکیبات حاصل از روش همبستگی توانست مقدار خطای برآورد دبی جریان را کاهش دهد که این نتیجه با نتایج حاصل از تحقیق حاضر مطابقت دارد. شکل‌های ۳ و ۴، به‌ترتیب نمودار مقادیر مشاهده‌ای و مقادیر برآورد شده SSL برای ترکیبات بهینه‌ی متغیر حاصل از آزمون گاما (GT) و روش آزمون و خطا با استفاده از مدل GEP را نشان می‌دهد.



شکل ۳. نمودار مقادیر مشاهده‌ای و برآورد شده SSL برای ترکیبات بهینه‌ی متغیر GT با استفاده از مدل GEP



شکل ۴. نمودار مقادیر مشاهده‌ای و برآورد شده SSL برای ترکیبات متغیر از روش آزمون و خطا با استفاده از مدل GEP

همان‌طور که در شکل‌های ۳ و ۴ مشاهده می‌شود، مدل ۱ برای ترکیبات بهینه‌ی متغیر حاصل از GT، با بهترین خط برازش توانست تا مقدار بار رسوب معلق را در مقادیر کم، متوسط و زیاد، با بالاترین دقت و صحت (R^2=۰/۹۹) برآورد کند.

نتیجه‌گیری

در این تحقیق، از مدل هوشمند برنامه‌ریزی بیان ژن برای برآورد بار رسوب معلق روزانه در حوزه‌ی آبخیز رود زرد در شرق استان خوزستان استفاده شد. بهترین ترکیبات از متغیرهای ورودی با استفاده از روش آزمون گاما به‌دست آمد و ترکیبات حاصل از این روش، با ترکیبات حاصل از روش آزمون و خطا، در طی فرآیند مدل‌سازی مورد مقایسه قرار گرفتند. نتایج نشان داد ترکیبات حاصل از روش GT، برآورد صحیح‌تر و دقیق‌تر از بار رسوب معلق در مقایسه با مدل‌های حاصل از ترکیبات متغیری با روش آزمون و خطا را دارد. همچنین، بهینه‌ترین ترکیب به‌دست آمده از روش آزمون گاما با متغیرهای متوسط دبی روزانه به همراه دو گام تأخیر زمانی و متوسط بارندگی روزانه به همراه سه گام تأخیر زمانی، کارآمدترین مدل برای برآورد دقیق بار رسوب

معلق با کمترین مقدار آماره‌های RMSE و MAE و بیشترین مقدار آماره‌های R^2 و NSE بود. این ترکیب کمترین مقدار آماره‌ی خطای استاندارد، آماره‌ی گاما و آماره‌ی V_{ratio} را در مقایسه با سایر ترکیبات متغیری حاصل از GT دارا بود. بنابراین، نتیجه‌گیری کلی از این تحقیق نشان داد که استفاده از روش آزمون گاما به‌عنوان یک روش پیش‌پردازش داده‌ها، با انتخاب ترکیباتی از متغیرهای ورودی مناسب، باعث کاهش خطای برآورد و افزایش تشابه مقادیر داده‌های مشاهداتی با داده‌های محاسباتی شده و در نتیجه عملکرد مدل GEP در برآورد بار رسوب معلق را افزایش می‌دهد.

منابع مورد استفاده

Angabini, S., Ahmadi, H., Feizni, S., Motamed Vaziri, B. and Ershadi, S. 2014. Suspended Sediment Concentration Estimation using Artificial Neural Networks and Fuzzy Rule Base Model Case Study: Jagin Dam. Journal of Applied Sciences Research. 10(14):12-17.

Azamathulla, H. 2013. Gene-expression programming to predict friction factor for Southern Italian Rivers. Neural Computing and Applications. 23:1421–1426.

Bagatur, T. and Onen, F. 2014. A predictive model on air entrainment by plunging water jets using GEP and ANN. KSCE Journal of Civil Engineering. 18(1): 304–314.

Barzegari, F., Yosefi, M. and Talebi, A. 2015. Estimating suspended sediment by Artificial Neural Network (ANN), Decision Trees (DT) and Sediment Rating Curve (SRC) models (Case study: Lorestan Province, Iran). Civil Engineering Infrastructures Journal. 48(2): 373-380.

Demirci, M. and Baltaci, A. 2013. Prediction of suspended sediment in river using fuzzy logic and multilinear regression approaches. Neural Computing and Applications. 23(1): 145-151.

Emamgholizadeh, S. and Karimi Demneh, R. 2018. The comparison of artificial intelligence models for the estimation of daily suspended sediment load: a case study on Telar and Kasilian Rivers in Iran. Water Science and Technology: Water Supply. 1-14. https://doi.org/10.2166/ws.2018.062.

Emamgholizadeh, S., Bateni, S.M., Shahsavani, D., Ashrafi, T. and Ghorbani, H. 2015. Estimation of soil cation exchange capacity using Genetic Expression Programming (GEP) and Multivariate Adaptive Regression Splines (MARS). Journal of Hydrology. 529:1590-1600.

Ferreira, C. 2001. Gene Expression Programming: A New Adaptive Algorithm for Solving Problems. Complex Systems. 13 (2): 87-129.

Galelli, S., Humphrey, G.B., Maier, H.R., Castelletti, A., Dandy, G.C. and Gibbs, M.S. 2014. An evaluation framework for input variable selection algorithms for environmental data-driven models. Environmental Modelling and Software. 62:33-51.

Guven, A., Talu, N.E. 2010. Gene Expression Programing for Estimating Suspended Sediment Yield in Middle Euphrates Basin, Turkey. Clean – Soil Air Water. 38(12):1159–1168.

Jajarmizadeh, M., Kakaei Lafdani, E., Harun, S. and Ahmadi, A. 2015. Application of SVM and SWAT models for monthly streamflow prediction, a case study in south of Iran. KSCE Journal of Civil Engineering. 19(1):345-357.

Jamalizadeh, M.R., Moghaddamnia, A., Piri, J., Arbabi, V., Homayounifar, M. and Shahryari, A. 2008. Dust stormprediction using ANNs techniques (a case study: Zabol city). World Academy of Science, Engineering and Technology. 43:512–520.

Kisi, O. and Ozkan, C. 2017. A new approach for modeling sediment-discharge relationship: Local weighted linear regression. Water Resources Management. 30(2):1-23.

Malik, A., Kumar, A. and Piri, J. 2017. Daily suspended sediment concentration simulation using hydrological data of Pranhita River Basin, India. Computers and Electronics in Agriculture. 138: 20–28.

Melesse, A.M., Ahmad, S., McClain, M.E., Wang, X. and Lim, Y.H. 2011. Suspended sediment load prediction of river systems: An artificial neural network. Agricultural Water Management. 98(5):855-866.

Muzzammil, M., Alama, J. and Danish, M. 2015. Scour prediction at bridge piers in cohesive bed using Gene Expression Programming. Aquatic Procedia. 4:789-796.

Noori, R., Karbassi, A. and Sabahi, M.S. 2009. Evaluation of PCA and gamma test techniques on ANN operation for weekly solid waste prediction. Journal of Environmental Management. 91:767-771.

Ouillon, S. 2018. Why and how do we study sediment transport? Focus on coastal zones and ongoing methods. Water. 10(4), 390 pp.

Remesan, R., Shamim, M.A. and Han, D. 2008. Model data selection using gamma test for daily solar radiation estimation. Hydrological Processes. 22:4301-4309.

Rashidi, S., Vafakhah, M., Kakaei Lafdani, E. and Javadi, M.R. 2016. Evaluating the support vector machine for suspended sediment load forecasting based on gamma test. Arabian Journal of Geosciences. 9(11). http://dx.doi.org/10.1007/s12517-016-2601-9.

Shamim, M.A., Hassan, M., Ahmad, S. and Zeeshan, M. 2016. A comparison of Artificial Neural Networks (ANN) and Local Linear Regression (LLR) techniques for predicting monthly reservoir levels. KSCE Journal of Civil Engineering. 20(2): 971–977.

Stefansson, A., Koncar, N. and Jones, A.J. 1997. A note on the Gamma test. Neural Computing and Applications. 5:131–133.

Wan Jaafar, W.Z., Liu, J. and Han, A. 2011. Input variable selection for median flood regionalization. Water Resources Research. 47:1-18.

Wu, W., Dandy, G. and Maier, H. 2014. Protocol for developing ANN models and its application to the assessment of the quality of the ANN model development process in drinking water quality modeling. Environmental Modeling and Software. 54:108-127.

ISSN 2251-7480

Evaluation of genetic expression programming model for suspended sediment load estimation based on data preprocessing using gamma test method (case study: Rood Zard Watershed)

Adele Alijanpour Shalmani[*1], Ali Reza Vaezi[2], and Mahmood Reza Tabatabaei[3]

1) Ph.D. Student, Department of Soil Science, University of Zanjan, Zanjan, Iran.
[*] Corresponding author: adele.alijanpour@gmail.com
2) Professor, Department of Soil Science, University of Zanjan, Zanjan, Iran.
3) Assistant Professor, Soil Conservation and Watershed Management Research Institute Agricultural Research, Education and Extension Organization (AREEO), Tehran, Iran.

Received: 11-11-2018 Accepted: 29-04-2019

Abstract

One of the effective factors to identify the problems of watersheds is the daily suspended sediment load. Due to the lack of sufficient data in direct measurement of daily suspended sediment, intelligent models like the Genetic Expression Programing model (GEP) can be used to estimate it. In this research, the data of the machine hydrometric station was used in the Rood Zard watershed with a statistical period of 36 years (1977-2012). The input variables of the GEP model include instantaneous flow discharge (Q), average daily flow discharge (Q_i) and average daily precipitation (P_i) with three steps of time delay and output variable to the model includes daily suspended sediment load. In order to reduce time and cost, pre-processing of input data into the GEP model was obtained using gamma test method and entered the GEP model along with non-preprocessing combinations of the test and error method. The results of comparison between all models showed that the best combination of input variable from gamma test with the lowest standard error is zero, gamma statistic is 0.000092 and V_{ratio} statistic is 0.012 and the combination of variables including average daily flow discharge with two steps of time delay and average daily precipitation with three steps of time delay, had the most accurate and correct estimate for suspended sediment load. This model had the lowest value of RMSE=1671.90 (ton/day) and MAE=475.68 (ton/day) and the highest value of R^2=0.99 and NSE=0.99 compared to other models. Therefore, the use of gamma test method as a data preprocessing method, by selecting combinations of appropriate input variables to models, an average of up to 40% of the estimated error (RMSE) of daily suspended sediment load compared to the inputs from the test and reduce the error and increase the performance of the GEP model in estimating the suspended sediment load by increasing the similarity between the values of observational data with computational data.

Keywords: Gamma test; Genetic Expression Programming; Sediment; Watershed