Water and Soil Management and Modeling

Online ISSN: 2783 - 2546

University of Mohaghegh Ardabili

Comparing the performance of the multiple linear regression classic method and modern data mining methods in annual rainfall modeling (Case study: Ahvaz city)

Pouya Allahverdipour [1], Mohammad Taghi Sattari [2]*

[1] M.Sc. Student, Department of Water Engineering, Faculty of Agriculture, Tabriz University, Tabriz, Iran
[2] Associate Professor, Department of Water Engineering, Faculty of Agriculture, Tabriz University, Tabriz, Iran

Abstract

Introduction

Prediction of hydrological variables, especially precipitation, is very important in the management and planning of water resources. For this reason, accurate estimation methods have always been of interest to researchers. Furthermore, due to the water crisis in different regions, it is necessary to use different methods to predict the rainfall and the resulting runoff so that comprehensive and appropriate management can be applied in the field of water distribution. Since the past, various methods have been developed and used by researchers to predict hydrological variables. The use of classical methods such as multiple linear regression to predict hydrological variables, especially precipitation, has been one of the most important and widely used methods that have had good results. Recently, data mining methods have been developed for this purpose. In this research, a comparison between the performance of the classic multiple linear regression and modern data mining methods was made in the annual rainfall modeling of Ahvaz city, and finally the best model in terms of performance was determined.

Materials and Methods

In this study, the annual rainfall of Ahvaz city has been investigated and modeled. Meteorological data from Ahvaz station was collected over a period of 30 years (1992-2021). The data validation tests including tests of homogeneity, normality, trend, and outlier data were performed. Annual rainfall modeling of Ahvaz city was done with Multiple Linear Regression (MLR), Principal Component Analysis (PCA), Gene Expression Programming (GEP), and Support Vector Machine (SVM). Finally, using the coefficient of determination (R^2), Root Mean Square of Errors (RMSE), Nash-Sutcliffe Efficiency (NSE), and Willmott index (WI), the accuracy and performance of the models were compared.

Results and Discussion

In this study, XLSTAT software was used to model rainfall with multiple linear regression. In order to simulate precipitation through the SVM model, it is possible to examine the types of kernel function, among which linear and polynomial kernels of the second and third degree, which are common types used in hydrology, are selected and through trial and error the optimal results of this The type of kernels was calculated. According to these results, the support vector machine model with third degree polynomial kernel was determined as the optimal method of precipitation modeling. In simulating the precipitation process using gene expression programming, because this model has the ability to select more effective variables and eliminate variables with less influence, therefore, in this project, all eight input factors are used to determine meaningful variables and for further investigation, in addition to the set The default mathematical operators of the program (F1), modes based on the values of the four main operators (F2) and the set of operators F3 and F4 have been used.

The results of the validation tests that check the homogeneity, trend, normality, and outlier data showed the good quality of the recorded data and the possibility of using them with a high percentage of confidence to continue the study. The results of comparing the models showed that the methods of PCA and GEP with R^2=0.85, NSE=0.85, and WI=0.96 and very little difference in RMSE equal 35.49 and 35.70, respectively. They have

Water and Soil Management and Modeling

Online ISSN: 2783 - 2546

University of Mohaghegh Ardabili

predicted the annual rainfall of Ahvaz with better performance and more accuracy compared to other models. Considering the water crisis in different regions of the country, especially in Ahvaz, it is suggested to use the methods introduced in this research to predict rainfall and runoff resulting from it, so that a comprehensive and appropriate management can be applied in the field of water distribution.

Conclusion

In this research, a comparison was made between classical statistical methods and some modern data mining methods in forecasting the annual rainfall of Ahvaz city. The hydrological data of Ahvaz synoptic meteorological station was collected in a period of 30 years (1371-1400) and first the data was verified using homogeneity, trend, normality and outlier data tests. The results showed the good quality of the recorded data and the possibility of using them with a high percentage of confidence. Multiple linear regression (MLR), principal component analysis (PCA), gene expression programming (GEP) and support vector machine (SVM) methods were used to model precipitation. The results of running the models were compared using the coefficient of explanation (R2), root mean square errors (RMSE), Nash-Sutcliffe efficiency (NSE) and Wilmot index (WI). The results showed that the methods of principal component analysis and gene expression programming with R2 criteria equal to 0.85, NSE equal to 0.85 and WI equal to 0.96 and a very small difference in RMSE values equal to 35.49 and 35.70, respectively, compared to Other models have better performance and more accuracy.

According to the results of this research, it is suggested to use modern data mining methods in addition to classical statistical methods in future researches. Also, it is necessary to pay attention to the use of functions and optimal factors of models to achieve the best results in future researches. Considering the water crisis in different parts of the country, especially in Ahvaz, it is suggested to use the methods introduced in this research to predict the rainfall and runoff caused by it, so that a comprehensive and appropriate management can be applied in the field of water distribution.

Keywords: Data mining, Gene Expression Programming, Principal Component Analysis, Regression, Support Vector

Article Type: Case Study

*Corresponding Author, E-mail: mtsattar@tabrizu.ac.ir

Citation: Allahverdipour, P., & Sattari, M.T. (2023). Comparing the performance of the multiple linear regression classic method and modern data mining methods in annual rainfall modeling (Case study: Ahvaz city). *Water and Soil Management and Modeling*, 3(2), 125-142.
DOI: 10.22098/mmws.2022.11337.1120
DOR: 20.1001.1.27832546.1402.3.2.9.3

Received: 23 August 2022, Received in revised form: 10 September 2022, Accepted: 13 September 2022, Published online: 13 September 2022
Water and Soil Management and Modeling, Year 2023, Vol. 3, No. 2, pp. 125-142
Publisher: University of Mohaghegh Ardabili

مدل‌سازی و مدیریت آب و خاک

شاپا الکترونیکی: ۲۷۸۳-۲۵٤٦

مقایسهٔ عملکرد روش کلاسیک رگرسیون خطی چندگانه و روش‌های داده‌کاوی نوین در مدل‌سازی بارش سالانه (مطالعهٔ موردی: شهر اهواز)

پویا اللهویردی‌پور[۱]، محمدتقی ستاری[۲*]

[۱] دانشجوی کارشناسی ارشد، گروه مهندسی آب، دانشکدهٔ کشاورزی، دانشگاه تبریز، تبریز، ایران

[۲] دانشیار، گروه مهندسی آب، دانشکدهٔ کشاورزی، دانشگاه تبریز، تبریز، ایران

چکیده

پیش‌بینی متغیرهای هیدرولوژیکی به‌ویژه بارش اهمیت بسیار زیادی در مدیریت و برنامه‌ریزی منابع آبی داشته و به همین دلیل روش‌هایی که بتوانند برآوردی دقیق از آن داشته باشند همواره مورد توجه پژوهش‌گران بوده است. در این پژوهش مقایسه‌ای بین عملکرد روش کلاسیک رگرسیون خطی چندگانه و روش‌های داده‌کاوی نوین در مدل‌سازی بارش سالانهٔ شهر اهواز انجام شده است. داده‌های هیدرولوژیکی مربوط به ایستگاه هواشناسی همدیدی اهواز در دورهٔ زمانی ۳۰ ساله (۱۳۷۱–۱۴۰۰) گردآوری شده و نسبت به کنترل کیفی داده‌ها با استفاده از آزمون‌های همگنی، روند، بهنجاری و ارزیابی داده‌های پرت اقدام شد. سپس جهت مدل‌سازی بارش از روش‌های رگرسیون خطی چندگانه (MLR)، تحلیل مؤلفه‌های اصلی (PCA)، برنامه‌نویسی بیان ژن (GEP) و ماشین بردار پشتیبان (SVM) استفاده شد. از ۷۰ درصد داده‌ها جهت آموزش و از ۳۰ درصد داده‌ها جهت صحت‌سنجی مدل‌ها استفاده شده و نتایج حاصل از اجرای مدل‌ها با استفاده از معیارهای ضریب تبیین (R^2)، جذر میانگین مربعات خطاها (RMSE)، راندمان نش–ساتکلیف (NSE) و شاخص ویلموت (WI) مقایسه شدند. نتایج نشان داد که روش‌های تحلیل مؤلفه‌های اصلی و برنامه‌نویسی بیان ژن با معیار R^2 برابر ۰/۸۵ و NSE برابر ۰/۸۵ و WI برابر ۰/۹۶ و اختلاف بسیار ناچیز در مقادیر RMSE به‌ترتیب برابر با ۳۵/۴۹ و ۳۵/۷۰ نسبت به سایر مدل‌ها عملکرد بهتر و دقت بیش‌تر در پیش‌بینی بارش سالانهٔ اهواز دارند. با توجه به بحران آب در نقاط مختلف کشور و به‌ویژه اهواز پیشنهاد می‌شود با استفاده از روش‌های معرفی شده در این پژوهش نسبت به پیش‌بینی بارش‌ها و رواناب‌های ناشی از آن اقدام شود تا مدیریت جامع و مناسبی در زمینهٔ توزیع آب اعمال شود.

واژه‌های کلیدی: بردار پشتیبان، برنامه‌نویسی بیان ژن، تحلیل مؤلفه‌های اصلی، داده‌کاوی، رگرسیون

نوع مقاله: مطالعهٔ موردی

*مسئول مکاتبات، پست الکترونیکی: mtsattar@tabrizu.ac.ir

استناد: اللهویردی پور، پویا، و ستاری، محمدتقی (۱۴۰۲). مقایسهٔ عملکرد روش کلاسیک رگرسیون خطی چندگانه و روش‌های داده‌کاوی نوین در مدل‌سازی بارش سالانه (مطالعهٔ موردی: شهر اهواز). *مدل‌سازی و مدیریت آب و خاک*، ۳(۲)، ۱۲۵–۱۴۲.

DOI: 10.22098/mmws.2022.11337.1120

DOR: 20.1001.1.27832546.1402.3.2.9.3

تاریخ دریافت: ۱۴۰۱/۰۶/۰۱، تاریخ بازنگری: ۱۴۰۱/۰۶/۱۹، تاریخ پذیرش: ۱۴۰۱/۰۶/۲۲، تاریخ انتشار: ۱۴۰۱/۰۶/۲۲

مدل‌سازی و مدیریت آب و خاک، سال ۱۴۰۲، دوره ۳، شماره ۲، صفحه ۱۲۵ تا ۱۴۲

ناشر: دانشگاه محقق اردبیلی

۱- مقدمه

کاهش اثرات بارش ناگهانی و شدید با انجام اقدامات ایمنی پیش‌گیرانه می‌تواند بسیار مفید باشد، به همین دلیل پیش‌بینی بارندگی اهمیت بسیار بالایی دارد. با توجه به تغییرات آب و هوایی، پیش‌بینی دقیق بارندگی پیچیده‌تر از قبل شده است. روش‌های داده‌کاوی می‌توانند بارندگی را از طریق استخراج الگوهای پنهان در میان ویژگی‌های آب و هوای داده‌های گذشته پیش‌بینی کنند. نیاز به اطلاعات وسیع، پیچیده بودن مدل‌های فیزیکی و نبود قطعیت ذاتی این فرآیند از دلایلی است که باعث شده پژوهش‌گران به سوی مدل‌های نوین روی آورند (Aftab et al., 2018). چندین روش مختلف برای پیش‌بینی بارش به‌کار گرفته می‌شوند که این روش‌ها به دو رویکرد دینامیکی و تجربی دسته‌بندی می‌شوند. رویکرد دینامیکی بارندگی را از طریق مدل‌های مبتنی بر فیزیک پیش‌بینی می‌کند. اصولاً از روش‌های عددی پیش‌بینی بارندگی در رویکردهای دینامیکی استفاده می‌شود. رویکرد تجربی به تجزیه و تحلیل سوابق بارندگی گذشته و ارتباط آن با سایر عوامل آب و هواشناسی می‌پردازد. روش‌هایی مانند رگرسیون، منطق فازی، شبکهٔ عصبی مصنوعی، برنامه‌نویسی بیان ژن و غیره از آن جمله‌اند. در میان آن‌ها، رگرسیون خطی چندگانه[1] (MLR)، به‌طور گسترده‌ای استفاده می‌شود، زیرا به‌راحتی رابطهٔ متغیرهای تأثیرگذار مختلف در به وجود آمدن بارندگی را به‌شکل معنی‌دار از نظر آماری نشان می‌دهد. هم‌چنین، به تعیین تاثیر یک متغیر بر سایر متغیرها کمک می‌کند (Andrews, 1974; Amiri et al., 2015).

در پژوهشی، Zaw and Naing (2008) یک مدل پیش‌بینی برای بارش میانمار با استفاده از رگرسیون خطی چندگانه ایجاد کردند. نتایج پژوهش آن‌ها نشان داد که مقادیر بارش پیش‌بینی شده با مقادیر واقعی مطابقت دارد. Dutta and Tahbilder (2014) بارش ناحیهٔ گواهاتی در هند را با استفاده از رگرسیون خطی چندگانه پیش‌بینی کردند. دماهای حداکثر و حداقل، رطوبت و فشار تراز دریا به‌عنوان عوامل پیش‌بینی‌کننده استفاده شدند. نتایج نشان داد که مدل پیش‌بینی بر اساس رگرسیون خطی چندگانه دارای دقت ۶۳ درصدی در پیش‌بینی بارش است. Swain et al. (2017) از مدل رگرسیون خطی چندگانه برای پیش‌بینی بارش سالانه در ناحیهٔ کاتک، در اودیشای هند استفاده کردند. پیش‌بینی‌های مدل ارتباط بسیار خوبی با داده‌های مشاهداتی را نشان داد، به‌طوری‌که ضریب تبیین (R^2) برابر ۰/۹۷ به‌دست آمد.

برای اولین بار Steiner (1965) بر اساس ۱۶ متغیر دما و رطوبت ماهانه، اقلیم امریکا را با استفاده از روش تحلیل مؤلفه‌های اصلی[2] (PCA) طبقه‌بندی کرد. Willmott (1978) با استفاده از روش تحلیل مؤلفه‌های اصلی مناطق بارشی ایالت کالیفرنیا در امریکا را ناحیه‌بندی کرد. هم‌چنین، Whetton (1988) تغییرات بارش در جنوب شرقی استرالیا را با تحلیل مؤلفه‌های اصلی بررسی کرد. Sneyers et al. (1989) نیز به تحلیل مؤلفه‌های اصلی بارندگی در بلژیک پرداختند. این بررسی نشان داد که چهار مؤلفهٔ اصلی ۹۹/۸ درصد از کل تغییرات را توضیح می‌دهند. Balafoutis (1991) از روش تحلیل مؤلفه‌های اصلی برای تعیین ویژگی‌های بارش کشور آلبانی استفاده کرد. داده‌های میانگین بارندگی ماهانه برای یک دورهٔ ۳۵ ساله (۱۹۶۵-۱۹۳۱) از ۱۵۴ ایستگاه هواشناسی تهیه شده‌اند. نتایج نشان داد که دو مؤلفهٔ اصلی اول ۸۷/۲ درصد از کل تغییرات را توصیف می‌کنند.

Baeriswyl and Rebetez (1997) به طبقه‌بندی مناطق کشور سوئیس از نظر بارش با استفاده از تحلیل مؤلفه‌های اصلی پرداختند. این روش امکان کاهش تعداد متغیرها از ۴۷ به ۵ را فراهم کرد. Stathis and Myronidis (2009) بارش در منطقه تسالی (یونان مرکزی) را به روش تحلیل مؤلفه‌های اصلی بررسی کردند. نتایج آن‌ها نشان داد که دو مؤلفهٔ اول ۸۷/۷ درصد از کل تغییرات را توضیح می‌دهند. Asakareh and Bayat (2013) به تحلیل مؤلفه‌های اصلی بارش سالانهٔ شهر زنجان پرداختند. نتایج نشان داد که با چهار مؤلفهٔ اول می‌توان ۹۵ درصد از تغییرات بارش سالانه را توضیح داد. در مطالعه‌ای Tripathi et al. (2006) از مدل ماشین بردار پشتیبان[3] (SVM) و شبکه‌های عصبی مصنوعی[4] (ANN) برای برآورد بارندگی ماهانه در هند استفاده کردند. مقایسهٔ نتایج برای چهار منطقهٔ معین هواشناسی هند در فصل مرطوب با معیار میانگین مربعات خطای نرمال شده[5] (NMSE) نشان داد که مدل ماشین بردار پشتیبان برای چهار منطقهٔ Bihar Plateau، Kerala، Coastal Karnataka و Orissa در مرحلهٔ آزمون به‌ترتیب دارای NMSEهای ۰/۴۶، ۰/۴۶، ۰/۹۶ و ۰/۲۷ بود درحالی که این معیار برای مدل ANN به‌ترتیب ۰/۴۹، ۰/۵۹، ۱/۵۱ و ۰/۷۳ بود. بنابراین، مدل ماشین بردار پشتیبان به‌دلیل مقدار NMSE کم‌تر در هر چهار منطقه، به عنوان گزینهٔ مناسب‌تر برای پیش‌بینی بارش تعیین شد. Lu and Wang (2011) در پژوهش‌های خود نشان دادند که ماشین بردار پشتیبان می‌تواند پیش‌بینی بارش را با میزان خطای کم انجام

[1] Multiple Linear Regression

[2] Principal Component Analysis
[3] Support Vector Machine
[4] Support Vector Machine
[5] Normalized Mean Square Error

دهد. مدل آن‌ها حدود ۹۹ درصد پیش‌بینی دقیق داشت. Hasan et al. (2015) به پیش‌بینی بارندگی بنگلادش با استفاده از رگرسیون بردار پشتیبان پرداختند. روش پیشنهادی دقت ۹۹/۹۲ درصدی در پیش‌بینی بارش داشت. در مطالعه‌ای Sureh et al. (2019) بارش ماهانهٔ ایستگاه چابهار را با استفاده از رگرسیون بردار پشتیبان و روش‌های نزدیک‌ترین همسایگی برآورد کردند. نتایج نشان داد که بهترین حالت روش رگرسیون بردار پشتیبان با استفاده از تابع هستهٔ چندجمله‌ای بهنجار شده است که معیار RMSE آن برابر ۱۴/۵۸ به‌دست آمد درحالی‌که این معیار برای روش نزدیک‌ترین همسایگی در بهترین حالت ۲۳/۱۸ بود. بنابراین روش رگرسیون بردار پشتیبان از دقت بالاتر و خطای تخمین کم‌تری برخوردار است.

Danandeh Mehr (2018) به پیش‌بینی بارندگی ماه‌های بعد در دو ایستگاه باران‌سنجی تبریز و ارومیه واقع در حوضهٔ دریاچهٔ ارومیه با استفاده از برنامه‌نویسی ژنتیک[1] (GP)، برنامه‌نویسی بیان ژن[2] (GEP) و مدل خودرگرسیون حالت–فضا[3] (ASS) پرداخت. معیار RMSE برای ایستگاه تبریز با مدل‌های فوق به‌ترتیب ۱۴۴/۸، ۱۲۵ و ۱۷۱ و برای ایستگاه ارومیه به‌ترتیب ۲۴۲، ۲۰۳ و ۲۶۳ به‌دست آمد. بنابراین، در هر دو ایستگاه بررسی شده، مدل تکامل یافته GEP از GP بهتر عمل کرده و به‌طور قابل‌توجهی نسبت به مدل‌های ASS برتری داشت. Solgi et al. (2018) از دو مدل هوشمند برنامه‌نویسی بیان ژن و ماشین بردار رگرسیونی برای پیش‌بینی بارش ماهانه شهرستان نهاوند استفاده کردند. نتایج به‌دست آمده نشان داد که عملکرد هر دو مدل با ضریب همبستگی برابر ۰/۹۲ خوب و مشابه بوده، ولی با توجه به معیار RMSE که مقدار آن برای مدل برنامه‌نویسی بیان ژن ۰/۰۴۷۸ و برای مدل ماشین بردار رگرسیونی ۰/۰۴۸۶ بود، مدل برنامه‌نویسی بیان ژن عملکردی کمی بهتر داشته است. Mirabbasi et al. (2019) از پنج روش هوش مصنوعی شامل برنامه‌نویسی بیان ژن، ماشین بردار پشتیبان، شبکهٔ عصبی مصنوعی، اسپلاین رگرسیونی تطبیقی چندمتغیره (MARS) و مدل درختی M5 برای پیش‌بینی بارش‌های درازمدت ایالت مادهایا پرادش هندوستان استفاده کردند. نتایج نشان داد که عملکرد مدل‌های مورد مطالعه نزدیک به یک‌دیگر بوده و از توانایی لازم برای پیش‌بینی عامل بارش برخوردارند.

با توجه به موارد بیان شده لزوم انجام یک پژوهش جهت بررسی و مقایسه‌ی کاربرد روش‌های آماری کلاسیک و روش‌های داده‌کاوی نوین در پیش‌بینی بارش سالانه احساس می‌شود. بنابراین، هدف از این پژوهش، بررسی نتایج حاصل از روش‌های رگرسیون خطی چندگانه (MLR)، تحلیل مؤلفه‌های اصلی (PCA)، برنامه‌نویسی بیان ژن (GEP) و ماشین بردار پشتیبان (SVM) در مدل‌سازی بارش سالانهٔ اهواز است.

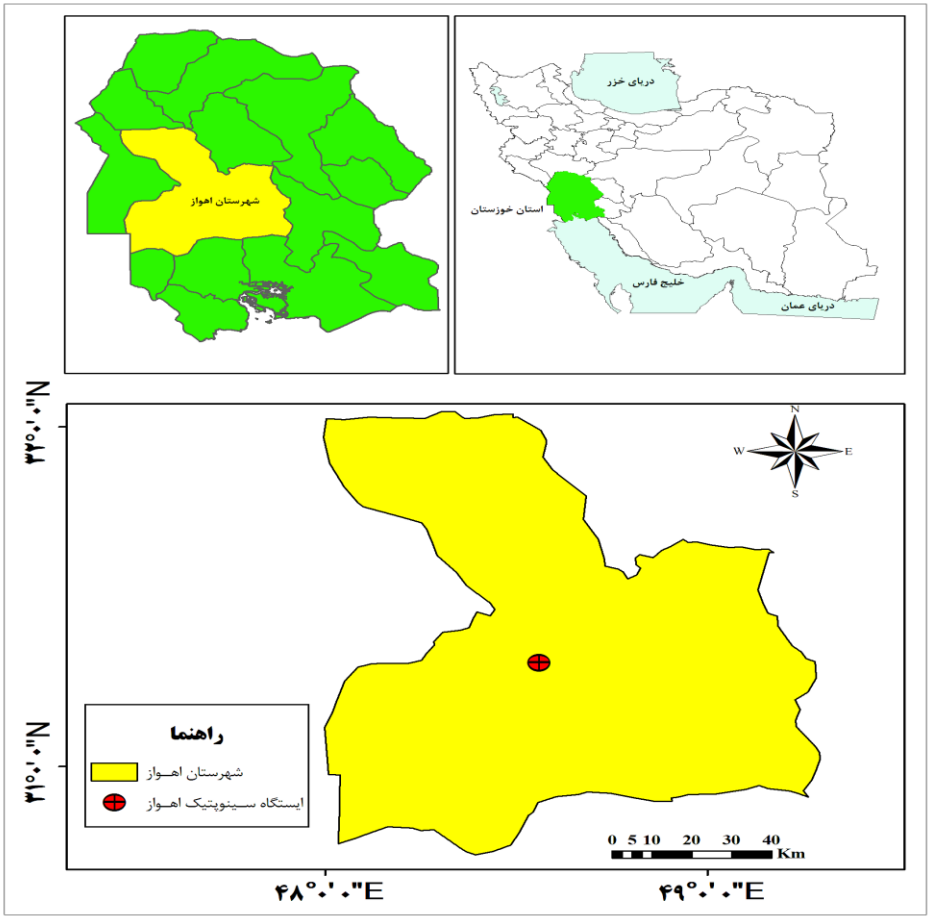
۲- مواد و روش‌ها

۲-۱- منطقه مورد مطالعه و داده‌های مورد استفاده

اهواز یکی از کلان‌شهرهای ایران است که در بخش مرکزی شهرستان اهواز قرار دارد و به‌عنوان مرکز استان خوزستان شناخته می‌شود. هم‌چنین، اهواز بزرگ‌ترین و پرجمعیت‌ترین شهر جنوب‌غربی ایران است. بخش بزرگی از استان خوزستان، جلگه است و شهر اهواز نیز در بخش جلگه‌ای جای دارد. میانگین دمای سالانهٔ شهر اهواز ۲۶/۳۷ درجهٔ سانتی‌گراد بوده و میانگین و انحراف معیار بارش سالانهٔ این شهر به‌ترتیب ۱۹۷/۱۹ و ۹۱/۱۱ میلی‌متر است. دمای هوای این شهر در زمستان تا پنج درجهٔ سانتی‌گراد کاهش و در تابستان تا ۵۰ درجهٔ سانتی‌گراد افزایش می‌یابد. ایستگاه هواشناسی همدیدی اهواز در عرض جغرافیایی ۳۱ درجه و ۲۰ دقیقه و طول جغرافیایی ۴۸ درجه و ۴۰ دقیقه و در ارتفاع ۲۲/۵ متری از سطح دریا واقع شده است. شکل ۱ موقعیت شهر اهواز در استان خوزستان و ایران و ایستگاه هواشناسی همدیدی اهواز را نشان می‌دهد.

در این پژوهش برای تحلیل و مدل‌سازی بارش اهواز داده‌های ایستگاه هواشناسی همدیدی اهواز در دورهٔ آماری ۳۰ ساله (۱۴۰۰–۱۳۷۱) از سایت www.tutiempo.net گردآوری شد. آموزش مدل‌ها با استفاده از ۷۰ درصد داده‌ها و صحت‌سنجی نیز با استفاده از ۳۰ درصد داده‌ها انجام شد. داده‌های مورد استفاده شامل بارش (pp)، دمای کمینه (T_m)، دمای بیشینه (T_M)، دمای میانگین (T)، میانگین سرعت باد افقی (V)، روزهای بارانی (RN)، روزهای برفی (SN)، روزهای طوفانی (TS) و روزهای مه‌آلود (FG) هستند. در جدول ۱ ویژگی‌های آماری شامل کمینه، بیشینه، میانه، میانگین و انحراف معیار داده‌های ۳۰ ساله ایستگاه اهواز آمده است.

[1] Genetic Programming
[2] Gene Expression Programming
[3] Autoregressive State-Space



شکل ۱- موقعیت ایستگاه هواشناسی همدیدی اهواز در استان خوزستان و ایران

Figure 1- location of Ahvaz synoptic meteorological station in Khuzestan province and Iran

جدول ۱- خلاصه ویژگی‌های آماری داده‌ها

Table 1- Summary statistics of data

آماره	pp (mm)	T (°c)	TM (°c)	Tm (°c)	V (km/hr)	RA (day/year)	SN (day/year)	TS (day/year)	FG (day/year)
حداقل	37.58	24.00	30.30	18.20	7.38	19.00	0.00	5.00	1.00
حداکثر	418.62	27.83	35.77	21.17	10.68	64.00	3.00	35.00	21.00
میانه	195.88	26.40	33.53	19.30	8.50	39.50	0.00	16.00	10.00
میانگین	197.19	26.37	33.44	19.38	8.70	38.63	0.30	17.43	10.57
انحراف معیار	91.11	0.76	1.18	0.56	0.88	9.57	0.64	7.13	5.41

جدول ۲ ماتریس همبستگی داده‌ها را نشان می‌دهد. مطابق این جدول مقدار بارش با تعداد روزهای مه‌آلود (FG) بیش‌ترین همبستگی مستقیم (۰/۲۲) و با تعداد روزهای برفی (SN) بیش‌ترین همبستگی معکوس (۰/۱۳-) را داشته است. بنابراین در سال‌های با تعداد روزهای مه‌آلود بیش‌تر، انتظار بارش بیش‌تر وجود دارد. هم‌چنین در سال‌های با تعداد روزهای برفی بیش‌تر در مقایسه با سال‌های با تعداد روزهای برفی کم‌تر، کم‌ترین میزان بارش انتظار می‌رود.

۲-۲- آماده‌سازی داده‌ها

لازمهٔ هر مطالعهٔ هیدرولوژی و منابع آب که در آن از سری‌های زمانی داده‌های هواشناسی و هیدرولوژی استفاده می‌شود، صحت‌سنجی داده‌ها است. واضح است که بدون اطمینان از صحت و کیفیت داده‌ها، نمی‌توان از آن‌ها برای استخراج نتایج بعدی استفاده کرد. به این‌منظور، پژوهش‌گران مختلف از آزمون‌های صحت‌سنجی متفاوت نظیر آزمون‌های همگنی[1] و ارزیابی داده‌های پرت[2] استفاده می‌کنند. در این آزمون‌ها روند

[1] Homogeneity Test
[2] Outlier data

تغییرات، بهنجاری[1]و استاندارد بودن داده‌ها و به‌طور کلی کیفیت و سلامت سری زمانی داده‌های تاریخی مورد ارزیابی و آزمون قرار می‌گیرند. بدین‌ترتیب، انجام آزمون‌های آماری مذکور، یک بخش جدایی‌ناپذیر در آغاز پژوهش‌های هیدرولوژی و منابع آب است (Ghajarnia et al., 2015).

۲-۳- رگرسیون خطی چندگانه (MLR)

رگرسیون یک روش آماری است که از رابطهٔ بین دو یا چند متغیر از میان داده‌های مشاهداتی برای پیش‌بینی متغیرهای دیگر استفاده می‌کند. انواع مختلفی از تحلیل رگرسیون وجود دارد که از بین آن‌ها رگرسیون خطی بسیار استفاده می‌شود زیرا استفاده از آن بسیار ساده است (Preacher et al., 2006; Nolan et al., 2015). در این روش فرض می‌شود که متغیرهای کمی به‌صورت خطی با یک‌دیگر مرتبط هستند. اساساً دو نوع رگرسیون خطی وجود دارد، رگرسیون خطی ساده و رگرسیون خطی چندگانه (Piña-Monarrez and Ortiz-Yañez, 2015). معادلهٔ رگرسیون خطی چندگانه که در این مطالعه استفاده شده، در رابطهٔ ۱ نشان داده شده است.

$$P = a + bX_1 + cX_2 + dX_3 + \cdots \quad (۱)$$

در یک معادلهٔ رگرسیون خطی چندگانه معمولی، یک متغیر باید تخمین زده شود که به‌عنوان متغیر پیش‌بینی (متغیر وابسته) درنظر گرفته می‌شود. متغیرهایی که پیش‌بینی از آن‌ها برآورد می‌شود به‌عنوان پیش‌بینی کننده (متغیر مستقل) در نظر گرفته می‌شوند. ضرایب رگرسیون (a, b, c, d, …) بر مقدار متغیر وابسته تاثیر دارند. با توجه به رابطهٔ (۱) می‌توان به‌وضوح متوجه شد که P متغیر وابسته است، در حالی‌که متغیرهای X_1, X_2, X_3, ... متغیرهای مستقل (پیش‌بینی‌کننده) هستند (Krzywinski and Altman, 2015). روش‌های مختلفی برای تعریف خطا و حداقل کردن آن وجود دارد. معیاری که در مدل رگرسیون خطی به‌کار می‌رود، کمینه کردن مجموع مربعات خطا است. از آن‌جایی که میانگین مقدارهای خطا صفر در نظر گرفته شده است، می‌دانیم زمانی مجموع مربعات خطا، حداقل ممکن را خواهد داشت که توزیع داده‌ها بهنجار باشند. در نتیجه، بهنجار بودن داده‌های متغیر وابسته یا باقی‌مانده‌ها یکی از فرضیات مهم برای مدل رگرسیونی خطی است.

۲-۴- روش تحلیل مؤلفه‌های اصلی (PCA)

تحلیل مؤلفه‌های اصلی (PCA) یک روش پرکاربرد است که مجموعهٔ بزرگی از متغیرها را به مجموعه‌ای کوچک‌تر و قابل تفسیرتر کاهش می‌دهد (Wilks, 2011). تحلیل مؤلفه‌های اصلی روشی آماری است که غالباً برای بررسی گروهی از متغیرهای همبسته به‌کار می‌رود. این روش به‌خصوص در شرایطی که ابعاد داده‌ها و ترکیب ساختار آن‌ها کاملاً مشخص نیست، مفید است. در این روش متغیرهای موجود در یک فضای چند حالته همبسته به یک مجموعه از مؤلفه‌های غیرهمبسته خلاصه می‌شوند که هریک از آن‌ها ترکیب خطی از متغیرهای اصلی هستند. مؤلفه‌های غیرهمبسته به‌دست‌آمده مؤلفه‌های اصلی نامیده می‌شوند که از بردارهای ویژه ماتریس کوواریانس یا ماتریس همبستگی متغیرهای اصلی به‌دست می‌آیند (Jolliffe and Cadima, 2016). مزیت اصلی کاربرد این روش از بین بردن هم‌خطی در مدل‌ها به واسطه تعداد زیاد متغیرهای مؤثر در مدل است.

به‌طورکلی کاربرد عمده روش تحلیل اجزای اصلی عبارت است از: کاهش تعداد متغیرها و یافتن ساختار ارتباطی بین متغیرها که در حقیقت همان دسته‌بندی متغیرها است. تعداد مؤلفه‌های استخراج شده در هر مدل برابر است با تعداد متغیرهایی که بررسی می‌شوند. اما می‌توان تعداد مشخصی از این مؤلفه‌ها را انتخاب کرد. معمولاً دو یا سه مؤلفهٔ اول مقدار قابل‌توجهی از پراکندگی داده‌ها را درنظر می‌گیرند. بنابراین، انتخاب دو یا سه مؤلفهٔ اول برای ادامه کار کفایت می‌کند (Jolliffe, 1993). اما در برخی از موارد ضروری است معیارهای دیگری را نیز برای یافتن تعداد مؤلفه‌های لازم مورد توجه قرارداد. این معیارها عبارتند از: آزمون اسکری[2]، معیار ویژه مقدارها[3]و معیار واریانس (Cattell, 1966).

۲-۵- ماشین بردار پشتیبان (SVM)

ماشین بردار پشتیبان مجموعه‌ای از روش‌های یادگیری تحت نظارت است که توسط Vapnik and Chervonenkis (2015) بر اساس نظریهٔ یادگیری آماری معرفی شده است. معمولاً برای طبقه‌بندی دوگانه در فضای ویژگی‌های دلخواه اجرا می‌شود؛ بنابراین، برای مسائل پیش‌بینی مناسب است (Pai and Hong, 2007). مدل‌های ماشین بردار پشتیبان به دو گروه اصلی تقسیم می‌شوند: یکی با موضوع طبقه‌بندی و دیگری پیش‌بینی و رگرسیون. روش دوم یک روش جدید مبتنی بر تئوری آماری است که از نگاشت غیرخطی برای آموزش داده‌ها استفاده می‌کند. در این حالت، الگوهای ورودی به فضایی با ابعاد بالا نگاشت

[1] Normality

[2] Scree Test

[3] Eigen Values

می‌شوند .دقت رگرسیون بردار پشتیبان به کمینه کردن تابع خطا مربوط می‌شود. ماشین‌های بردار پشتیبان ابعاد مسأله را از طریق توابع هسته برای حل مسائل غیرخطی تغییر می‌دهند. انتخاب هسته به اندازهٔ داده‌های آموزشی و ابعاد بردار ویژگی بستگی دارد. باتوجه به این عوامل باید یک تابع هسته انتخاب شود که قابلیت آموزش ورودی‌های مسئله را داشته باشد.

جدول ۲- ماتریس همبستگی داده‌ها
Table 2- Correlation matrix of data

	pp	T	TM	Tm	V	RA	SN	TS	FG
pp	1.00	0.11	0.10	0.19	-0.11	0.03	-0.13	0.19	0.22
T	0.11	1.00	0.93	0.59	-0.18	-0.05	-0.13	0.26	0.08
TM	0.10	0.93	1.00	0.33	-0.38	0.08	-0.16	0.38	0.09
Tm	0.19	0.59	0.33	1.00	0.14	-0.29	-0.18	-0.13	-0.12
V	-0.11	-0.18	-0.38	0.14	1.00	-0.42	0.28	-0.36	0.03
RA	0.03	-0.05	0.08	-0.29	-0.42	1.00	-0.05	0.76	0.53
SN	-0.13	-0.13	-0.16	-0.18	0.28	-0.05	1.00	-0.17	-0.01
TS	0.19	0.26	0.38	-0.13	-0.36	0.76	-0.17	1.00	0.57
FG	0.22	0.08	0.09	-0.12	0.03	0.53	-0.01	0.57	1.00

۲-۶- برنامه‌نویسی بیان ژن (GEP)

برنامه‌نویسی بیان ژن نوع خاصی از برنامه‌نویسی ژنتیک (GP) چندشاخه‌ای است که به ایجاد درخت‌های بیانی[1] (ETs؛ کروموزوم‌ها) حاوی یک یا چند ژن، که هر کدام یک درخت بیان فرعی[2]را کد می‌کند، اجازه می‌دهد تا یک مسأله خاص را حل کند. در GEP متغیر خروجی با پیوند دادن زیر ETهای مربوطه با استفاده از توابع جبری یا بولی (And، Or، Not) محاسبه می‌شود. در GP برنامه‌های کامپیوتری (ژن‌ها) از زبان LISP پیروی می‌کنند و به‌صورت درختان تجزیه با اندازه‌ها و شکل‌های مختلف بیان می‌شوند. در مقابل در GEP برنامه‌های کامپیوتری به‌عنوان رشته‌های خطی با طول ثابت متشکل از یک یا چند ژن در نظر گرفته می‌شوند. هر ژن شامل یک سر و دم است. برای تعداد معین ژن و طول سر، جمعیت اولیهٔ کروموزوم‌های GEP از طریق پر کردن دامنه‌های سر و دم با توجه به توالی کدگذاری ژن‌ها به نام چارچوب خوانش باز ایجاد می‌شوند. فرآیند تکامل در GEP زمانی خاتمه می‌یابد که الگوریتم GEP راه‌حل کامل را بیابد (برازش بهترین کروموزوم به مقدار مشخص شده برسد) یا تعداد ژن‌های تولید شده بدون توجه به کیفیت جواب به‌دست آمده به تعداد معینی رسیده باشد. در این روش پدیده‌های مختلف با استفاده از مجموعه‌ای از توابع و ترمینال‌ها، مدل‌سازی می‌شوند. مجموعه توابع، معمولاً شامل توابع حسابی، مثلثاتی و یا توابع تعریف شده توسط کاربر هستند که معتقد است، می‌توانند برای تفسیر مدل مناسب باشند. مجموعه ترمینال‌ها، از مقادیر ثابت و متغیرهای مستقل مسأله تشکیل شده‌اند. توضیحات بیش‌تر در مورد GEP از Ferreira (2002) قابل استفاده است.

۲-۷- معیارهای ارزیابی عملکرد مدل‌ها

به‌منظور ارزیابی دقت و میزان عملکرد مدل‌های بررسی شده در این پژوهش و مقایسه آن‌ها با بارش مشاهداتی از معیارهای ضریب تبیین (R^2)، جذر میانگین مربعات خطاها (RMSE)، راندمان نش–ساتکلیف (NSE) و شاخص ویلموت (WI) استفاده شد. رابطه‌های ۲ تا ۵ معیارهای آماری فوق را نشان می‌دهند.

$$R^2 = \frac{\left[\sum_{i=1}^{N}(O_i - \overline{O}_i)(P_i - \overline{P}_i)\right]^2}{\sum_{i=1}^{N}(O_i - \overline{O}_i)^2 \sum_{i=1}^{N}(P_i - \overline{P}_i)^2} \quad (۲)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(O_i - P_i)^2}{N}} \quad (۳)$$

$$NSE = 1 - \frac{\sum_{i=1}^{N}(O_i - P_i)^2}{(O_i - \overline{O}_i)^2} \quad (۴)$$

$$WI = 1 - \left[\frac{\sum_{i=1}^{N}(O_i - P_i)^2}{\sum_{i=1}^{N}(|P_i - \overline{O}_i| + |O_i - \overline{P}_i|)^2}\right] \quad (۵)$$

در رابطه‌های فوق O_i مقادیر مشاهداتی (واقعی)، P_i مقادیر پیش‌بینی شده، $\overline{O}_i$ میانگین مقادیر مشاهداتی، $\overline{P}_i$ میانگین مقادیر پیش‌بینی شده و N تعداد مشاهدات هستند. مقادیر R^2، NSE و WI بیش‌تر و هم‌چنین مقادیر RMSE کم‌تر، نشان‌دهندهٔ دقت بیش‌تر و عملکرد بهتر مدل هستند.

[1] Expression Trees
[2] Sub-ET

۳- نتایج و بحث

۳-۱- نتایج مربوط به آماده‌سازی داده‌ها

جهت آماده‌سازی اولیه داده‌ها و کنترل کیفی آن‌ها آزمون‌های همگنی، روند و بهنجاری انجام و داده‌های پرت ارزیابی شدند.

۳-۱-۱- ارزیابی همگنی

در این پژوهش به‌منظور ارزیابی همگنی داده‌ها از چهار آزمون شامل؛ آزمون همگنی نرمال استاندارد[1] (SNHT)، آزمون همگنی دامنه بیشند[2] آزمون همگنی پتیت[3] و آزمون نسبت فون‌نیومن[4] (VNR) استفاده شده است. این آزمون‌ها در نرم‌افزار آماری XLSTAT اجرا شده‌اند. در جدول ۳ نتایج مربوط به آزمون همگنی داده‌ها ارائه شده است. در این آزمون‌ها فرض صفر بیان‌گر همگن بودن داده‌ها و فرض یک بیان‌گر غیرهمگن بودن داده‌ها است. چنان‌که مقدار P-Value از مقدار درجهٔ اطمینان α بزرگ‌تر باشد، فرض صفر صحیح است و در غیر این صورت فرض یک قابل‌قبول است. با توجه به این‌که سه آزمون از چهار آزمون فوق، همگنی داده‌ها را تأیید کردند؛ بنابراین، این داده‌ها همگن هستند (P>۰/۰۵).

۳-۱-۲- ارزیابی روند

با استفاده از آزمون روند من-کندال[5] که یک روش ناپارامتریک است و هم‌چنین روش شیب سن[6] روند تغییرات داده‌ها ارزیابی شد. قابل ذکر است فرض صفر عدم وجود روند و فرض یک دارای روند بودن داده‌هاست. در جدول ۳ نتایج آزمون روند من-کندال آمده است (Sattari and Judi 2018; Kazemzadeh et al., 2019). چون مقدار P-Value بیش‌تر از مقدار درجهٔ اطمینان α به‌دست آمده، بنابراین فرض صفر یعنی عدم وجود روند در داده‌ها مورد قبول است (P>۰/۰۵). در منحنی شیب سن نیز خط روند تقریباً ثابت بوده و شیب آن حدوداً صفر است که عدم وجود روند را نشان می‌دهد (شکل ۲).

۳-۱-۳- ارزیابی بهنجاری داده‌ها

این آزمون‌ها به‌منظور ارزیابی تبعیت داده‌های مورد استفاده از توزیع نرمال انجام می‌شوند. در این پژوهش چهار آزمون Shapiro-Wilk، Anderson-Darling، Lilliefors و Jarque-Bera انجام شده است. قابل ذکر است فرض صفر بهنجاری و فرض یک بهنجار نبودن داده‌ها است. همان‌طور که در جدول ۳ مشاهده می‌شود، چون مقادیر P-Value در هر چهار آزمون فوق از مقدار درجهٔ اطمینان α کم‌تر هستند؛ بنابراین، فرض صفر یعنی بهنجاری داده‌ها رد می‌شود (P>۰/۰۵). بنابراین، باید جهت ادامه پژوهش نسبت به بهنجارسازی داده‌ها اقدام کرد. در این پژوهش جهت بهنجارسازی داده‌ها از روش تبدیل باکس کاکس[7] استفاده شد.

۳-۱-۴- ارزیابی داده‌های پرت

با استفاده از آزمون‌های Grubbs test و Dixon test داده‌های پرت ارزیابی شدند. فرض صفر عدم وجود داده پرت و فرض یک وجود داده پرت در بین کل داده‌ها است. طبق جدول ۳ چون در هر دو آزمون مقادیر P-Value از مقدار درجهٔ اطمینان α کم‌تر است؛ بنابراین، فرض یک؛ یعنی عدم وجود داده‌پرت رد شده و فرض یک تأیید می‌شود (P>۰/۰۵). در ادامه نسبت به تصحیح داده‌های پرت با توجه به اینکه فقط یک دادهٔ پرت وجود داشت، با حذف آن و درنظر گرفتن آن به‌عنوان دادهٔ گم‌شده اقدام شد. چندین روش برای پر کردن داده‌های گم‌شده وجود دارد. می‌توان آن ردیف داده‌ها را از فرآیند محاسبات حذف کرد، با میانگین کل داده‌ها در آن قسمت جایگزین کرد یا با توجه به اعداد قبل و بعد از آن، نسبت به پرکردن آن داده اقدام کرد. در این‌جا از میانگین داده‌ها استفاده شد.

۳-۲- مدل‌سازی بارش اهواز

پس از ارزیابی اولیه داده‌ها و کنترل کیفی آن‌ها، جهت مدل‌سازی بارش اهواز از روش‌های رگرسیون خطی چندگانه (MLR)، تحلیل مؤلفه‌های اصلی (PCA)، برنامه‌نویسی بیان ژن (GEP) و ماشین بردار پشتیبان (SVM) استفاده شده و سپس نتایج به‌دست آمده توسط مدل‌ها مقایسه شدند.

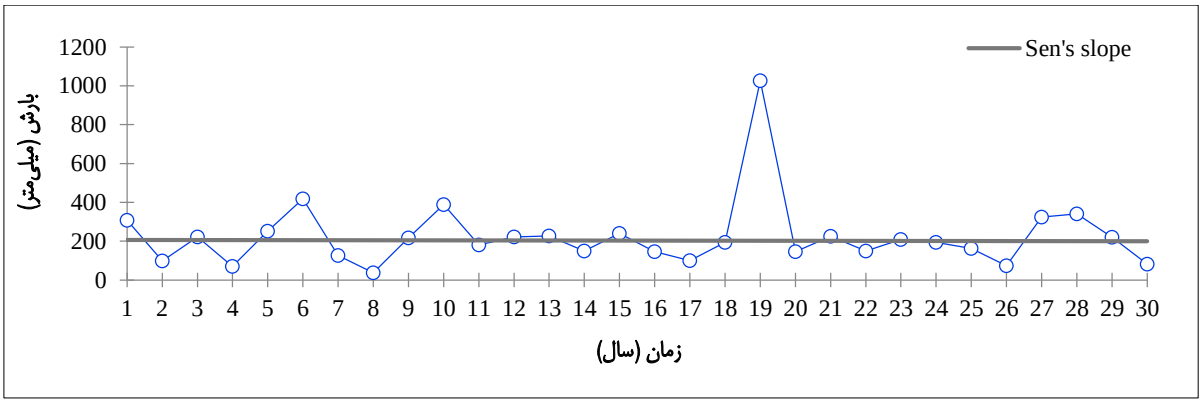
۳-۲-۱- رگرسیون خطی چندگانه (MLR)

در این پژوهش جهت مدل سازی بارش با رگرسیون چندگانه خطی از نرم‌افزار XLSTAT استفاده شده است. مدل‌سازی بارش بر اساس سایر عوامل هواشناسی و با استفاده از چهار روش انتخاب مدل که در نرم‌افزار XLSTAT وجود دارد، شامل بهترین

[1] Standard Normal Homogeneity Test
[2] Buishand Range Homogeneity Test
[3] Pettitt Homogeneity Test
[4] Von Neumann Ratio Test
[5] Mann-Kendall
[6] Sen's Slope
[7] Box-Cox Transformation

جدول ۳- نتایج آزمون‌های صحت‌سنجی داده‌ها
Table 3- Results of data validation tests

ردیف	نوع آزمون	نام آزمون	α	P-Value
1	همگنی	Pettitt	0.05	0.000
		Standard Normal	0.05	0.989
		Buishand	0.05	0.874
		Von Neumann Ratio	0.05	0.661
2	روند	Mann-Kendall	0.05	0.920
3	بهنجاری	Shapiro-Wilk	0.05	0.0001
		Anderson-Darling	0.05	0.0001
		Lilliefors	0.05	0.000
		Jarque-Bera	0.05	0.0001
4	داده‌های پرت	Grubbs	0.05	0.0001
		Dixon	0.05	0.0001



شکل ۲- منحنی شیب سن بارش مشاهداتی
Figure 2- Sen's slope curve of observed precipitation

مدل[1] گام به گام[2] پیش رو[3] و پس رو[4] استفاده شده و در نهایت بهترین روش انتخاب مدل رگرسیونی تعیین شد. در جدول ۴ مشاهده می‌شود که روش انتخاب مدل رگرسیونی با روش بهترین مدل دارای بیش‌ترین مقدار ضریب تبیین (R^2)، کم‌ترین مقدار جذر میانگین مربعات خطاها (RMSE)، بیش‌ترین مقدار راندمان نش–ساتکلیف (NSE) و بیش‌ترین مقدار شاخص ویلموت (WI) است؛ بنابراین، به‌عنوان روش انتخاب مدل رگرسیونی تعیین می‌شود.

در نهایت رابطهٔ رگرسیونی بارش با استفاده از روش انتخابی بهترین مدل به‌صورت رابطهٔ (۶) به‌دست آمد:

$$pp = 990.13 + 140.44 \times T - 113.03 \times T_M - 47.60 \times T_m + 6.19 \times V - 0.15 \times RA - 20.66 \times SN + 8.46 \times TS + 1.62 \times FG \quad (۶)$$

هم‌چنین، بهنجار بودن توزیع خطاها به روش Shapiro-Wilk ارزیابی شد. با توجه به جدول ۵ چون مقدار P-Value از مقدار درجهٔ اطمینان α بیش‌تر است؛ بنابراین، فرض صفر؛ یعنی بهنجار بودن توزیع خطاها تأیید می‌شود.

مطابق شکل ۳ که پراکنش مقادیر مشاهداتی و خطاهای استاندارد شده را نشان می‌دهد، روند ثابتی (افزایشی یا کاهشی) وجود ندارد. بنابراین فرض اولیه خطی بودن رابطه تأیید می‌شود که به معنی تأیید رابطهٔ رگرسیونی است.

نمودار پراکنش مقادیر پیش‌بینی شده توسط مدل رگرسیون خطی چندگانه (MLR) و مقادیر مشاهداتی در شکل ۴ رسم شده است. مطابق این نمودار با توجه به اینکه پراکنش مقادیر پیش‌بینی شده توسط مدل و مقادیر مشاهداتی در اطراف خط

[1] Best Model
[2] Stepwise
[3] Forward
[4] Backward

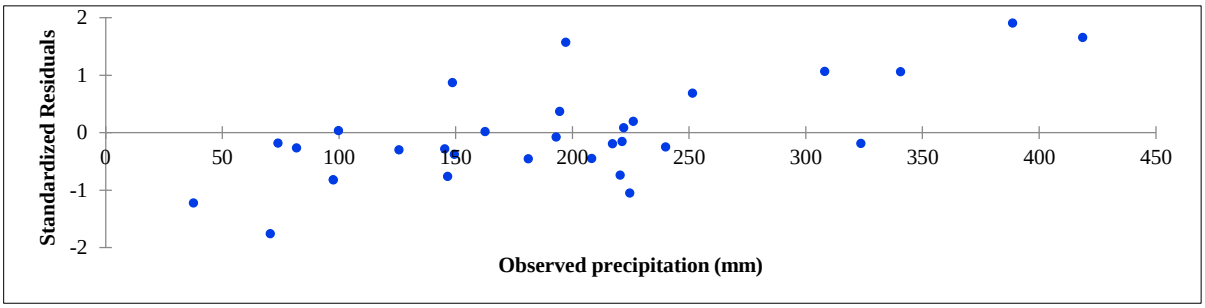
نیمساز و در فاصله تلورانس یا محدوده مورد نظر قرار دارند؛ بنابراین، استفاده از رگرسیون خطی در این پروژه قابل قبول است و بهنجاری توزیع مانده‌های رگرسیون تأیید می‌شود.

جدول ۴- نتایج ارزیابی مدل‌های رگرسیونی در پیش‌بینی بارش
Table 4-Evaluation results of regression models in predict of precipitation

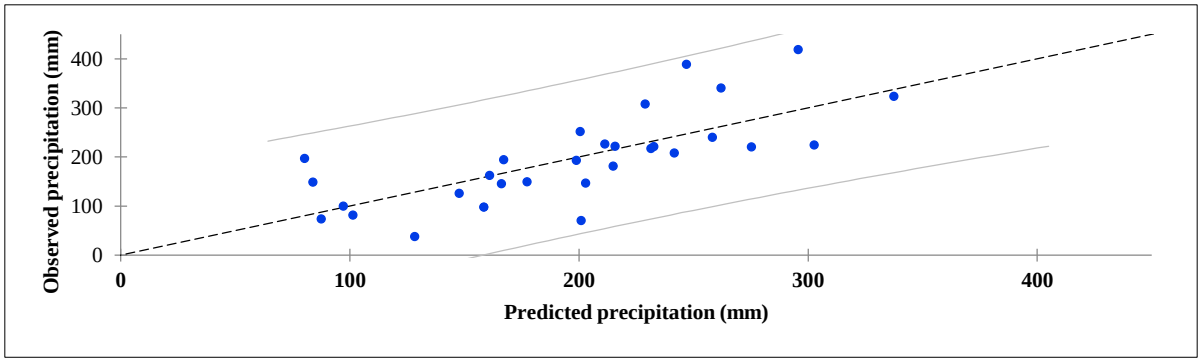
انتخاب بهترین مدل	R^2	RMSE	NSE	WI
Best model	0.535	62.11	0.54	0.82
Stepwise	0.434	68.55	0.43	0.76
Forward	0.434	68.55	0.43	0.76
Backward	0.434	68.55	0.43	0.76

جدول ۵- نتایج آزمون Shapiro-Wilk برای ارزیابی بهنجاری خطاهای روش رگرسیون خطی چندگانه
Table 5- Results of Shapiro-Wilk test to evaluate the normality of errors of the multiple linear regression method

P-Value	alpha
0.17	0.05



شکل ۳- نمودار پراکنش مقادیر مشاهداتی و خطاهای استاندارد شده مقادیر پیش‌بینی شده به روش رگرسیون خطی چندگانه
Figure 3- Scatter Plot of observed values and Standardized Residual of predicted values by multiple linear regression method



شکل ۴- نمودار پراکنش بارش پیش‌بینی شده با مدل رگرسیون خطی چندگانه و بارش مشاهداتی
Figure 4- Scatter plot of predicted precipitation by the Multiple Linear Regression (MLR) model and observed precipitation

۳-۲-۲- تحلیل مؤلفه‌های اصلی (PCA)

نتایج به‌دست آمده با روش تحلیل مؤلفه‌های اصلی با استفاده از نرم‌افزار XLSTAT در ادامه ارائه شده است. جدول ۶ که جدول ویژه مقدار نامیده می‌شود، میزان ارزش و تأثیر هر یک از مؤلفه‌ها یا عوامل (F_i) را نشان می‌دهد. مطابق ردیف دوم که مقدار ویژه عوامل است، می‌توان عوامل یک تا چهار را در مدل‌سازی استفاده‌کرد. طبق ردیف سوم که مقدار تجمعی مقدار ویژه عوامل را نشان می‌دهد، عوامل یک تا چهار حدود ۸۵ درصد تغییرات را پوشش می‌دهند.

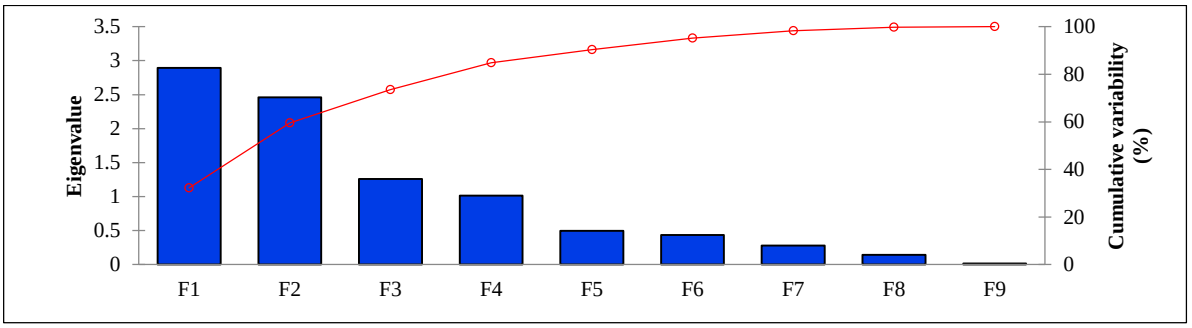
۳-۲-۲-۱- نمودار اسکری[1]

[1] Scree Plot

در نمـودار اسـکری مقادیر مؤلفه‌ها در مقابل مقدار ویژه آن‌ها رسم می‌شوند (شکل ۵). هم‌چنین، منحنی تجمعی آن‌ها نیز رسم می‌شوند که مجموعاً برابر ۱۰۰ درصد هستند. با توجه به محور سمت راست نمودار، که درصد تجمعی تاثیر عوامل را نشان می‌دهد، مشاهده می‌شود که می‌توان با عوامل یک تا چهار، حدود ۸۵ درصد از تغییرات را پوشش داد. بنابراین، این عوامل به‌عنوان مؤلفه‌های اصلی انتخاب می‌شوند.

جدول ۶- جدول ویژه مقدار
Table 6- Eigenvalue table

	F1	F2	F3	F4	F5	F6	F7	F8	F9
Eigenvalue	2.896	2.461	1.260	1.013	0.496	0.436	0.278	0.145	0.015
Variability (%)	32.175	27.348	14.002	11.250	5.514	4.844	3.087	1.614	0.166
Cumulative %	32.175	59.523	73.525	84.776	90.290	95.134	98.220	99.834	100



شکل ۵- نمـودار اسـکری مؤلفه‌های اصلی
Figure 5- Scree plot of principal components

در شکل ۶ نمودار مهمی که تحلیل داده‌ها و متغیرها را می‌توان از روی آن انجام داد، رسم شده است. این نمودار براساس دو عامله که ارزش ویژه بالاتری دارند (عوامل ۱ و ۲) رسم می‌شود. محورها از ۱- تا ۱+ هستند که مقدار همبستگی‌ها را نشان می‌دهند. یعنی از روی محور x میزان همبستگی متغیرها با عامل ۱ و از روی محور y هم میزان همبستگی متغیرها با عامل ۲ نمایش داده می‌شود. هرکدام از بردارها در نمودار نشان‌دهندهٔ متغیرها هستند و کوچک یا بزرگی هر بردار نشان‌دهندهٔ همبستگی آن متغیر با عوامل است. در شکل ۶ مشاهده می‌شود که به‌عنوان مثال عامل ۱ با متغیر TS بیش‌ترین همبستگی را دارد.

۳-۲-۲-۲- نمودار دو مختصه‌ای[1]

این نمودار هم‌زمان هم تاثیر متغیرها (به‌صورت بردار) و هم تاثیر تک تک مشاهدات[2](به‌صورت نقطه‌ای) را بر عوامل انتخاب شده نشان می‌دهد. با توجه به شکل ۷ مشاهده می‌شود که مثلاً Q بر عامل یک، متغیرهای FG، RA و TS و هم‌چنین مشاهدات ۲۷، ۲۸ و ۲۹ بیش‌ترین تأثیر را دارند.

با توجه به نتایج تحلیل مؤلفه‌های‌اصلی، از عوامل یک تا چهار که بیش‌ترین تأثیر را داشتند، جهت ایجاد رابطهٔ رگرسیونی استفاده شد. ابتدا نسبت به تعیین مقادیر این عوامل با توجه به جدول امتیاز عوامل و رابطهٔ (۷) اقدام می‌شود:

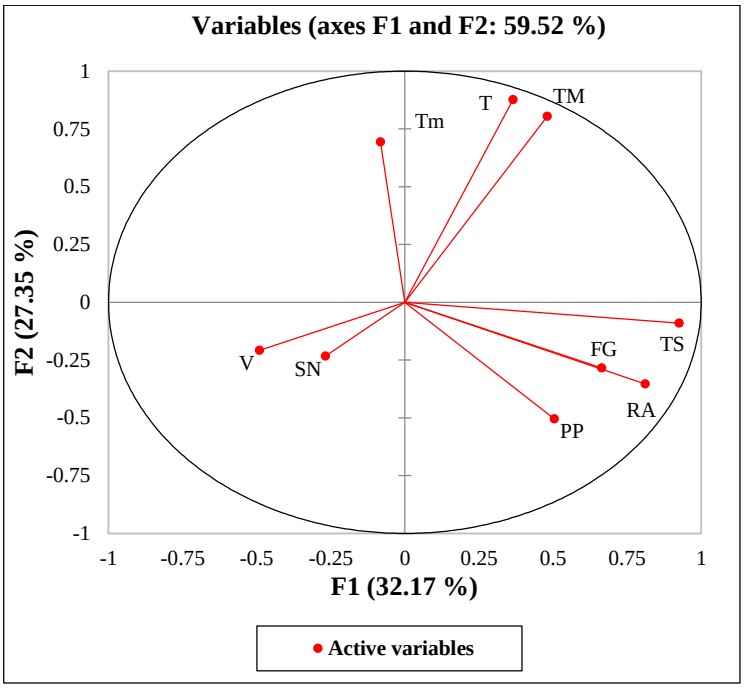
$$PC_i = \sum_{i=1}^{n} W_i X_i \qquad (۷)$$

در رابطهٔ ۷، PC_i مؤلفهٔ اصلی i ام، W_i تاثیر هر متغیر بر عامل و X_i هم متغیرها هستند. n نیز تعداد مشاهدات است. در نهایت، رابطهٔ رگرسیونی با مؤلفه‌های اصلی به‌صورت رابطهٔ (۸) تعیین شد:

$$pp = 179.46 + 0.16 \times F_1 - 0.086 \times F_2 + 0.086 \times F_3 - 0.081 \times F_4 \qquad (۸)$$

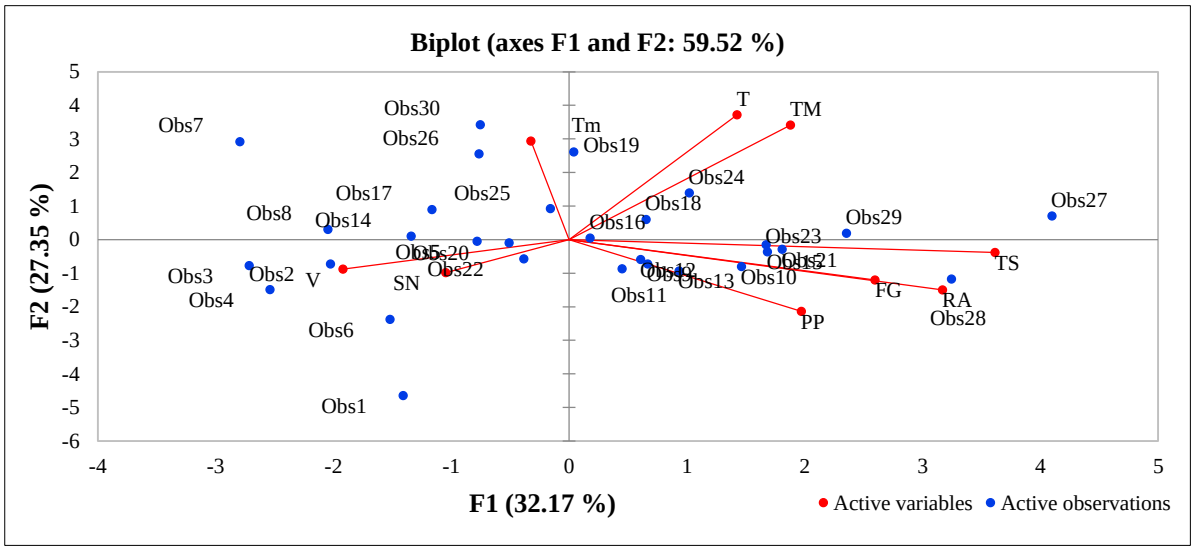
مطابق جدول ۷ روش تحلیل مؤلفه‌های اصلی قابل قبول ارزیابی شد. هم‌چنین بهنجاری توزیع خطاها (باقیمانده‌ها) به چهار روش ذکر شده در بالا ارزیابی شده و در نتیجه بهنجاری توزیع خطاها تأیید شد.

[1] Biplot
[2] Observations



شکل ۶– نمودار همبستگی عوامل اصلی با متغیرها

Figure 6- Correlation plot between Principal Component's with variables



شکل ۷– نمودار دو مختصه‌ای بارش مشاهداتی و متغیرهای مورد بررسی با دو عامل اصلی اول

Figure 7- Biplot of the observed precipitation and the studied variables with the first two Principal Components

جدول ۷– ارزیابی روش تحلیل مؤلفه‌های اصلی (PCA) در پیش‌بینی بارش

Table 7- Evaluation of the method of Principal Component Analysis (PCA) in predict of precipitation

تعداد متغیرها	متغیرها	R^2	RMSE	NSE	WI
4	F1 / F2 / F3 / F4	0.85	35.49	0.85	0.96

۳-۲-۳- ماشین بردار پشتیبان (SVM)

به‌منظور شبیه سازی بارش از طریق مدل SVM می‌توان انواع تابع کرنل را مورد بررسی قرارداد که از بین آن‌ها کرنل‌های خطی و چندجمله‌ای درجهٔ دو و درجهٔ سه، که از انواع رایج مورد استفاده در هیدرولوژی میباشند، انتخاب و از طریق سعی و خطا نتایج بهینه این نوع کرنل‌ها محاسبه شد که نتایج در جدول ۸ آمده است. با توجه به این نتایج مدل ماشین بردار پشتیبان با کرنل چندجمله‌ای درجهٔ سه به‌عنوان روش بهینه مدل‌سازی بارش تعیین شد.

جدول ۸- ارزیابی ساختارهای مدل ماشین بردار پشتیبان (SVM) در پیش‌بینی بارش

Table 8- Evaluation of Support Vector Machine (SVM) model structures in predict of precipitation

شماره	کرنل	معیار ارزیابی			
		WI	NSE	RMSE	R^2
1	خطی	0.72	0.37	72.17	0.40
2	چندجمله‌ای درجهٔ دو	0.82	0.52	62.93	0.54
3	چندجمله‌ای درجهٔ سه	0.93	0.79	41.47	0.82

۳-۲-۴- برنامه‌نویسی بیان ژن (GEP)

در شبیه‌سازی فرآیند بارش با استفاده از برنامه‌ریزی بیان ژن به‌دلیل این‌که این مدل توان انتخاب متغیرهای مؤثرتر و حذف متغیرهای با تأثیرگذاری کم‌تر را دارد، لذا در این پروژه از کل هشت عامل ورودی برای تعیین متغیرهای معنادار استفاده و برای بررسی بیش‌تر علاوه بر مجموعه عملگرهای ریاضی پیش فرض برنامه (F_1)، حالت‌هایی بر اساس مقادیر چهار عملگر اصلی (F_2) و مجموعه عملگرهای F_3 و F_4 استفاده شده است.

$$F_1 = \left\{ +, -, \times, \div, Exp, Ln, \frac{1}{x}, X^2, \min(x,y), \max(x,y), avg(x,y), \sqrt[3]{x}, Atan, Tanh, (1-x) \right\}$$

$$F_2 = \{+, -, \times, \div\}$$

$$F_3 = \{+, -, \times, \div, X^2\}$$

$$F_4 = \{+, -, \times, \div, X^3\}$$

عوامل مورد استفاده و نرخ آن‌ها در استخراج مدل‌های شبیه‌سازی برای فرآیند بارش با استفاده از روش برنامه‌نویسی بیان ژن که با استراتژی بهینه‌سازی از گزینه‌های پیش‌فرض نرم‌افزار انتخاب شد، در جدول ۹ ارائه شده است.

جدول ۹- مقادیر پارامترهای مورد استفاده در روش برنامه‌نویسی بیان ژن (GEP)

Table 9- Values of factors used in Gene Expression Programming (GEP model)

تنظیمات کلی		عملگرهای ژنتیکی	
تعداد کروموزوم‌ها	۳۰	نرخ جهش	۰/۰۰۱۳۸
اندازه راس	8	نرخ وارون‌سازی	۰/۰۰۵۴۶
تعداد ژن‌ها در هر کروموزوم	3	نرخ ترانهش درج متوالی	۰/۰۰۵۴۶
تابع پیوند	جمع (+)	نرخ ترکیب تک نقطه‌ایی	۰/۰۰۲۷۷
معیارخطا	RMSE	نرخ ترکیب دو نقطه‌ایی	۰/۰۰۲۷۷
		نرخ ترکیب ژن	۰/۰۰۲۷۷
		نرخ ترانهش ژن	۰/۰۰۲۷۷

نتایج ارزیابی اجرای مدل برنامه‌نویسی بیان ژن با عملگرهای به‌کار گرفته شده در جدول ۱۰ آمده است. در تحقیق حاضر عملگر F_4 با بیش‌ترین ضریب تبیین (R^2=0.85)، کم‌ترین ریشهٔ میانگین مربعات خطا (RMSE=35.70)، بیش‌ترین مقدار راندمان نش–ساتکلیف (NSE=0.85) و بیش‌ترن مقدار شاخص ویلموت (WI= 0.96) در مقایسه با سایر عملگرها از دقت بیش‌تری برخوردار است، درحالی‌که F_1 (مجموعه عملگرهای ریاضی پیش فرض برنامه) ضعیف‌ترین عملکرد را داشته است. بنابراین در تحقیقات باید بررسی شود که کدام مجموعه عملگر نتیجه بهینه این مدل را تولید می‌کند.

جدول ۱۰- ارزیابی مدل برنامه‌نویسی بیان ژن (GEP) در پیش‌بینی بارش

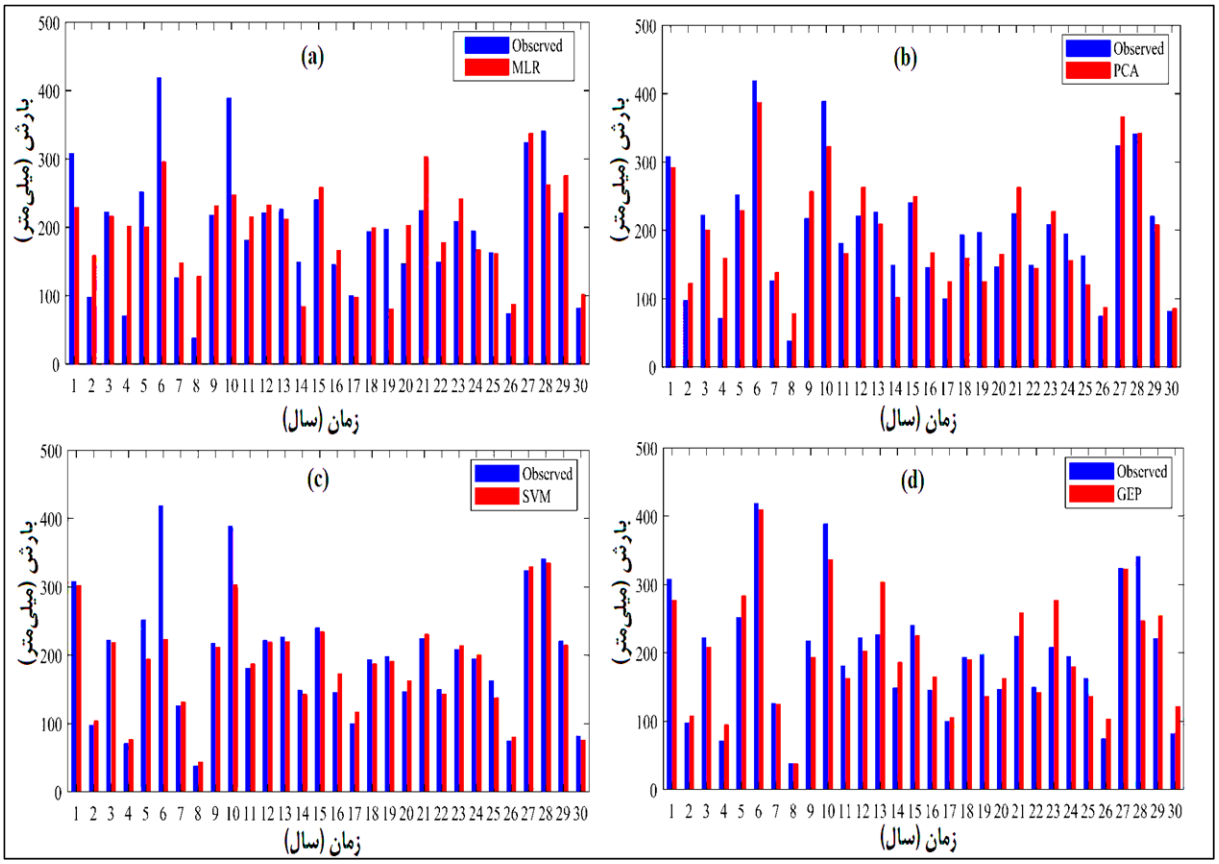
Table 10- Evaluation of Gene Expression Programming (GEP) model in predict of precipitation

شماره	عملگر	معیار ارزیابی			
		WI	NSE	RMSE	R^2
1	F_1	0.74	0.43	68.81	0.44
2	F_2	0.87	0.59	58.41	0.60
3	F_3	0.95	0.82	38.94	0.82
4	F_4	0.96	0.85	35.70	0.85

۳-۳- مقایسهٔ عملکرد مدل‌ها

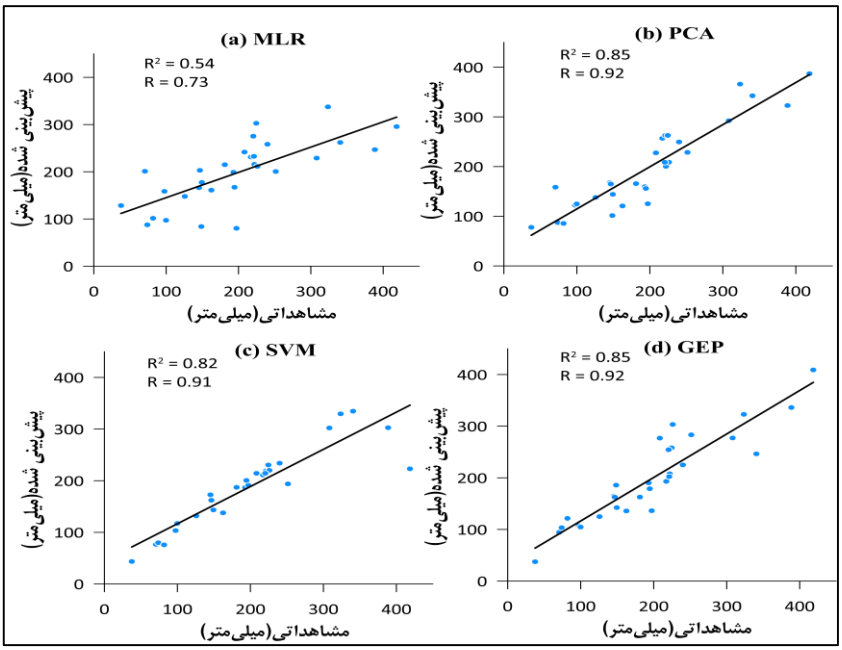
با بررسی مدل‌های ذکر شده و با انتخاب مقدار بهینه هر مدل مقادیر شبیه‌سازی بارش سالانهٔ اهواز به‌دست آمد. در شکل ۸ نمودار میله‌ای و در شکل ۹ نمودار پراکنش مقادیر مشاهداتی و پیش‌بینی شده توسط مدل‌ها رسم شده است. در جدول ۱۱ نیز نتایج مدل‌ها با توجه به معیارهای آماری ارزیابی شده‌اند. با توجه به شکل‌های ۸ و ۹ و مقایسهٔ معیارهای ارزیابی به‌دست آمده مطابق جدول ۱۱، مشاهده می‌شود که به‌ترتیب روش تحلیل مؤلفه‌های اساسی (PCA)، برنامه‌ریزی بیان ژن (GEP)، ماشین بردار پشتیبان (SVM) و رگرسیون خطی چندگانه (MLR) دقت بیش‌تر و عملکرد بهتری در برآورد بارش سالانهٔ اهواز داشته‌اند. نتایج با پژوهش مشابهی که Solgi et al. (2018) انجام داده بودند و از دو مدل هوشمند برنامه‌نویسی بیان ژن و ماشین بردار رگرسیونی برای پیش‌بینی بارش ماهانه شهرستان نهاوند استفاده کرده بودند، مطابقت دارد که مدل برنامه‌نویسی بیان ژن عملکرد کمی بهتری داشت. البته در این پژوهش بارش ماهانه بررسی شده بود که باتوجه به نتایج مشابه در استفاده از این مدل‌ها در پژوهش حاضر جهت مدل‌سازی بارش سالانه، می‌تواند نشان‌دهندهٔ کارایی بسیار مناسب این نوع مدل‌ها در هر دو مقیاس ماهانه و سالانه باشد. در پژوهش Mirabbasi et al. (2019) نیز که از پنج روش هوش مصنوعی شامل برنامه‌نویسی بیان ژن، ماشین بردار پشتیبان، شبکهٔ عصبی مصنوعی، MARS و مدل درختی M5 برای پیش‌بینی بارش‌های درازمدت ایالت مادهایا پرادش هندوستان استفاده شد نیز عملکرد مدل‌های مورد مطالعه نزدیک به یک‌دیگر بود که مشابه نتایج این پژوهش است. مشابهت نتایج این پژوهش می‌تواند به‌دلیل استفاده از داده‌های درازمدت در هر دو پژوهش باشد.

شکل ۱۰ دیاگرام تیلور را برای مدل‌های بررسی شده در این پژوهش نشان می‌دهد. مطابق این دیاگرام مدل‌های PCA و GEP دارای کم‌ترین RMSE و بیش‌ترین ضریب همبستگی هستند. هم‌چنین، همهٔ مدل‌ها انحراف معیار کم‌تری نسبت به مشاهدات داشته‌اند و دو مدل PCA و GEP نزدیک‌ترین مقدار به انحراف معیار مربوط به مشاهدات را دارند.



شکل ۸- نمودار میله‌ای بارش مشاهداتی و پیش‌بینی شده توسط مدل‌های (a) MLR، (b)PCA، (c) SVM و (d)GEP

Figure 8- Bar plot of Observed and predicted precipitation by (a) MLR, (b) PCA, (c) SVM, and (d) GEP models



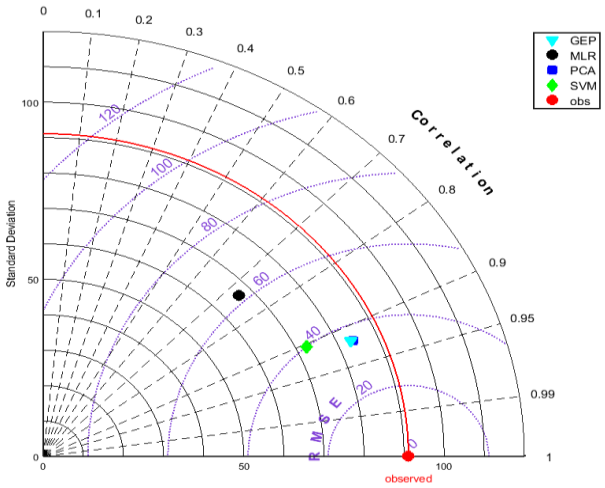
شکل ۹- پراکنش بارش مشاهداتی و پیش‌بینی شده توسط مدل‌های (a) MLR، (b) PCA، (c) SVM و (d) GEP

Figure 9- Scatter plot of observed and predicted precipitation by (a) MLR, (b) PCA, (c) SVM, and (d) GEP models

جدول ۱۱- نتایج ارزیابی مدل‌ها در پیش‌بینی بارش

Table 11-Results of models evaluation in predict of precipitation

شماره	مدل	معیار ارزیابی			
		WI	NSE	RMSE	R^2
1	رگرسیون خطی چندگانه (MLR)	0.82	0.54	62.11	0.54
2	تجزیه مؤلفه‌های اصلی (PCA)	0.96	0.85	35.49	0.85
3	ماشین بردار پشتیبان (SVM)	0.93	0.79	41.47	0.82
4	برنامه‌نویسی بیان ژن (GEP)	0.96	0.85	35.70	0.85



شکل ۱۰- دیاگرام تیلور برای ارزیابی عملکرد مدل‌ها در پیش‌بینی بارش

Figure 10- Taylor Diagram for evaluating the performance of models in predict of precipitation

۴- نتیجه‌گیری

در این پژوهش مقایسه‌ای میان روش‌های آماری کلاسیک و چند روش داده‌کاوی نوین در پیش‌بینی بارش سالانهٔ شهر اهواز انجام شد. داده‌های هیدرولوژیکی ایستگاه هواشناسی همدیدی اهواز در دورهٔ زمانی ۳۰ سال (۱۴۰۰-۱۳۷۱) گردآوری و ابتدا نسبت به صحت‌سنجی داده‌ها با استفاده از آزمون‌های همگنی، روند،

بهنجاری و بررسی داده‌های پرت اقدام شد. نتایج نشان‌دهندهٔ کیفیت خوب داده‌های ثبت شده و امکان استفاده از آن‌ها با درصد بالایی از اطمینان بود. جهت مدل‌سازی بارش از روش‌های رگرسیون خطی چندگانه (MLR)، تحلیل مؤلفه‌های اصلی (PCA)، برنامه‌نویسی بیان ژن (GEP) و ماشین بردار پشتیبان (SVM) استفاده شد. نتایج حاصل از اجرای مدل‌ها با استفاده از معیارهای ضریب تبیین (R^2)، جذر میانگین مربعات خطاها (RMSE)، راندمان نش–ساتکلیف (NSE) و شاخص ویلموت (WI) مقایسه شدند. نتایج نشان داد که روش‌های تحلیل مؤلفه‌های اصلی و برنامه‌نویسی بیان ژن با معیار R^2 برابر ۰/۸۵ و NSE برابر ۰/۸۵ و WI برابر ۰/۹۶ و اختلاف بسیار ناچیز در مقادیر RMSE به‌ترتیب برابر با ۳۵/۴۹ و ۳۵/۷۰ نسبت به سایر مدل‌ها عملکرد بهتر و دقت بیش‌تر داشته‌اند.

با توجه به نتایج این پژوهش پیشنهاد می‌شود در پژوهش‌های آینده از روش‌های داده‌کاوی نوین در کنار روش‌های آماری کلاسیک استفاده شود. هم‌چنین، توجه به استفاده از توابع و عوامل بهینه مدل‌ها جهت دست‌یابی به بهترین نتیجه در پژوهش‌های آتی ضروری است. با توجه به بحران آب در نقاط مختلف کشور و به‌ویژه اهواز پیشنهاد می‌شود با استفاده از روش‌های معرفی شده در این پژوهش نسبت به پیش‌بینی بارش‌ها و رواناب‌های ناشی از آن اقدام شود تا مدیریت جامع و مناسبی در زمینهٔ توزیع آب اعمال شود.

منابع

ستاری، محمدتقی، و رضازاده جودی، علی (۱۳۹۷). مدل‌سازی رواناب ماهانه با استفاده از روش‌های داده‌کاوی بر اساس الگوریتم‌های انتخاب ویژگی. *حفاظت منابع آب و خاک*، ۷(۴)، ۳۹–۵۴.

سلگی، اباذر، زارعی، حمید، شهنی دارابی، مهرنوش، و علیدادی ده کهنه، صابر (۱۳۹۷). کاربرد مدل‌های برنامه‌ریزی بیان ژن و ماشین بردار رگرسیونی جهت مدل‌سازی و پیش‌بینی بارش ماهانه. *تحقیقات کاربردی علوم جغرافیایی*، ۱۸(۵۰)، ۹۱–۱۰۳. doi:10.29252/jgs.18.50.91

عساکره، حسین، و بیات، علی (۱۳۹۲). تحلیل روند و چرخه ویژگی‌های بارندگی سالانهٔ زنجان. *جغرافیا و برنامه‌ریزی*، ۱۷(۴۵)، ۱۲۱–۱۴۲. doi:10.1016/j.atmosres.2015.02.010

کاظم‌زاده، مجید، ملکیان، آرش، مقدم‌نیا، علیرضا، و خلیقی، شهرام (۱۳۹۸). ارزیابی اثرات تغییر اقلیم بر خصوصیات هیدرولوژیکی حوزه آبخیز (مطالعه موردی: حوزه آبخیز آجی چای). *علوم و مهندسی آبخیزداری ایران*، ۱۳(۴۵)، ۱–۱۱. dor:20.1001.1.20089554.1398.13.45.1.5

References

Aftab, S., Ahmad, M., Hameed, N., Bashir, M.S., Ali, I., & Nawaz, Z. (2018). Rainfall prediction in Lahore City using data mining techniques. *International Journal of Advanced Computer Science and Applications*, 9(4), 254-260. doi:10.1016/j.enbuild.2015.09.073

Amiri, S.S., Mottahedi, M., & Asadi, S. (2015). Using multiple regression analysis to develop energy consumption indicators for commercial buildings in the US. *Energy and Buildings*, 109, 209-216. doi:10.2307/1267603

Andrews, D.F. (1974). A robust method for multiple linear regression. *Technometrics*, *16*(4), 523-531.

Asakareh, H., & Bayat, A. (2013). The analysis of the trend and the cycles of annual precipitation characteristics of Zanjan. *Geography and Planning*, 17(45), 121-142. [In Persian]

Baeriswyl, P.A., & Rebetez, M. (1997). Regionalization of precipitation in Switzerland by means of principal component analysis. *Theoretical and Applied Climatology*, 58(1), 31-41. doi.org/10.1007/BF00867430

Balafoutis, C.J. (1991). *Principal component analysis of Albanian rainfall* (No. RefW-15-14613). Aristotle University of Thessaloniki.

Cattell, R.B. (1966). The Scree test for the number of factors. *Multivariate Behavioral Research*, 1(2), 245-276. doi:10.1207/s15327906mbr0102_10

Danandeh Mehr, A.D. (2018). Month ahead rainfall forecasting using gene expression programming. *American Journal of Earth and Environmental Sciences*, 1(2), 63-70.

Dutta, P.S., & Tahbilder, H. (2014). Prediction of rainfall using data mining technique over Assam. *Indian Journal of Computer Science and Engineering (IJCSE)*, 5(2), 85-90.

Ferreira, C. (2002). Gene expression programming in problem solving. Pp. 635-653, In: *Soft computing and industry*, Springer, London.

Ghajarnia, N., Liaghat, A., & Arasteh, P.D. (2015). Comparison and evaluation of high-resolution precipitation estimation products in Urmia Basin-Iran. *Journal of Water and Soil Resources Conservation*, 4(1), 91-109. doi:10.1016/j.atmosres.2015.02.010

Hasan, N., Nath, N.C., & Rasel, R.I. (2015). A support vector regression model for forecasting rainfall. 2nd International Conference on Electrical Information and Communication Technologies (EICT), Pp. 554-559.

Jolliffe, I.T. (1993). Principal component analysis: a beginner's guide-II. Pitfalls, myths and extensions. *Weather*, 48(8), 246-253. doi:10.1002/j.1477-8696.1993.tb05899.x

Jolliffe, I.T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065), 20150202. doi:10.1098/rsta.2015.0202

Kazemzadeh, M., Malekian, A., Moghaddamnia, A. R., & Sigaroudi, K. (2019). Evaluation of Climate Change Impacts on Hydrological Characteristics of Watershed (Case study: Aji-Chai Watershed). *Iranian Journal of Watershed Management Science and Engineering*, 13(45), 1-11. dor:20.1001.1.20089554.1398.13.45.1.5 [In Persian]

Krzywinski, M., & Altman, N. (2015). Multiple linear regression. *Nature Methods*, 12(12), 1103-1104. doi:10.1038/nmeth.3665

Lu, K., & Wang, L. (2011). A novel nonlinear combination model based on support vector machine for rainfall prediction. 4th International Joint Conference on Computational Sciences and Optimization, Pp. 1343-1346.

Mirabbasi, R., Kisi, O., Sanikhani, H., & Gajbhiye Meshram, S. (2019). Monthly long-term rainfall estimation in Central India using M5Tree, MARS, LSSVR, ANN and GEP models. *Neural Computing and Applications*, 31(10), 6843-6862. doi:10.1007/s00521-018-3519-9

Nolan, B.T., Fienen, M.N., & Lorenz, D.L. (2015). A statistical learning framework for groundwater nitrate models of the Central Valley, California, USA. *Journal of Hydrology*, 531, 902-911. doi:10.1016/j.jhydrol.2015.10.025

Pai, P.F., & Hong, W.C. (2007). A recurrent support vector regression model in rainfall forecasting. *Hydrological Processes*, 21(6), 819-827. doi:10.1002/hyp.6323

Piña-Monarrez, M.R., & Ortiz-Yañez, J.F. (2015). Weibull and lognormal Taguchi analysis using multiple linear regression. *Reliability Engineering & System Safety*, 144, 244-253. doi:10.1016/j.ress.2015.08.004

Preacher, K.J., Curran, P.J., & Bauer, D.J. (2006). Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics*, 31(4), 437-448. doi:10.3102/10769986031004437

Sattari, M. T. & Rezazadeh Judi, A. (2018). Monthly runoff modeling using data mining methods based on feature selection algorithms. *Protection of Water and Soil Resources*, 7(4), 39-54. [In Persian]

Solgi, A., Zarei, H., Shahni, D.M., & Alidadi, D.K.S. (2018). Application of gene expression programming and support vector regression models to modeling and prediction monthly precipitation. *Journal of Geographical Sciences*, 18(50), 91-103. doi:10.29252/jgs.18.50.91 [In Persian]

Sneyers, R., Vandiepenbeeck, M., & Vanlierde, R. (1989). Principal component analysis of Belgian rainfall. *Theoretical and applied Climatology*. 39(4), 199-204. doi:10.1007/BF00867948

Stathis, D., & Myronidis, D. (2009). Principal component analysis of precipitation in Thessaly region (Central Greece). *Global Network of Environmental Science and Technology Journal*, 11(4), 467-476.

Steiner, D. (1965). *A Multivariate Statistical Approach to Climatic Regionalization and Classification*. EJ Brill.

Sureh, F.S., Sattari, M.T., & İrvem, A. (2019). Estimation of monthly precipitation based on machine learning methods by using meteorological variables. *Mustafa Kemal Üniversitesi Tarım Bilimleri Dergisi*, 24, 149-154.

Swain, S., Patel, P., & Nandi, S. (2017). A multiple linear regression model for precipitation forecasting over Cuttack district, Odisha, India. The 2nd International Conference for Convergence in Technology, Pp. 355-357. doi:10.1109/I2CT.2017.8226150

Tripathi, S., Srinivas, V.V., & Nanjundiah, R.S. (2006). Downscaling of precipitation for climate change scenarios: a support vector machine approach. *Journal of Hydrology*, 330(3-4), 621-640. doi:10.1016/j.jhydrol.2006.04.030

Vapnik, V.N., & Chervonenkis, A.Y. (2015). On the uniform convergence of relative frequencies of events to their probabilities. Pp. 11-30, In: Vovk, V., Papadopoulos, H., Gammerman, A. (eds) Measures of Complexity. Springer, Cham. doi:10.1007/978-3-319-21852-6_3

Whetton, P.H. (1988). A synoptic climatological analysis of rainfall variability in southeastern Australia. *Journal of Climatology*, 8(2), 155-177. doi:10.1002/joc.3370080204

Wilks, D.S. (2011). *Statistical methods in the atmospheric sciences* (Vol. 100). Academic press.

Willmott, C.J. (1978). P-mode principal components analysis, grouping and precipitation regions in California. *Archives for Meteorology Geophysics and Bioclimatology Series B Theoretical and Applied Climatology*, 26(4), 277-295. doi:10.1007/BF02243232

Zaw, W.T., & Naing, T.T. (2008). Empirical statistical modeling of rainfall prediction over Myanmar. *International Journal of Computer and Information Engineering*, 2(10), 3418-3421. doi:10.5281/zenodo.1084254