

Analysis of the role of low-probability events on the accuracy of discharge estimation for different return periods (case study: Godarkhush river basin, Ilam)

Z. Akbari¹, A. Talebi^{2*}, M. R. Ekhtesasi³, N. Rostami⁴

Extended Abstract

Introduction:

Floods are among the natural disasters that pose a serious threat to lives and property. Flood frequency analysis is essential for estimating design discharges for hydraulic structures and managing flood risk. Two primary methods for estimating design floods include statistical and hydrological approaches. Statistical methods typically use probability distribution functions such as Gumbel, log-normal, and Pearson Type III to analyze peak discharges. In areas with insufficient recorded data, rainfall data and regional analyses are utilized. Low-probability events and sparse historical data significantly impact analysis results. These data may lead to deviations in discharge estimations and alter return period assessments. Various statistical methods have been introduced for identifying and removing outliers; however, incorporating historical data can enhance modeling accuracy. Studies have shown that including historical data in statistical analyses improves the precision of design flood estimates. This study aims to examine outlier data at the Takht Khan hydrometric station, located in the Godarkhush River Basin, Ilam, and analyze the impact of removing or retaining these data on the accuracy of discharge estimates for different return periods.

Methods:

This study investigates flood estimation in the Godarkhush River watershed in Ilam Province, Iran, using statistical and hydrological analysis methods. The watershed, covering approximately 810 km², consists of mountainous terrain with seasonal river discharge variations. Key meteorological and hydrometric data were collected, including 34 years of rainfall and discharge records. Outlier detection is crucial in hydrological studies as extreme values can significantly impact statistical analyses and flood prediction accuracy. This study employed three methods for identifying outliers: graphical methods (box plots and scatter plots), statistical methods (Z-score analysis), and model-based methods (cluster analysis). The box plot method visualizes data dispersion and identifies extreme values, while scatter plots reveal potential anomalies in variable relationships. The Z-score method standardizes data by measuring deviations from the mean, with values exceeding ± 3 considered outliers. Cluster analysis groups similar data points and detects anomalies based on deviations from main clusters. For flood frequency analysis across different return periods, the Log-Pearson Type III distribution was selected for discharge modeling based on the goodness-of-fit tests, including the Anderson-Darling and Chi-square tests. Given the natural skewness of flood data, this distribution provides more reliable estimates of peak discharge for various return periods. The method involves logarithmic transformation of the data, calculation of the mean, standard deviation, and skewness coefficient, followed by fitting the Pearson Type III distribution. Accurate flood estimation is essential for the design of hydraulic structures and effective flood risk management. Identifying and appropriately handling outlier data improves data quality and enhances

1 - PhD Candidate in Watershed Sciences and Engineering, Department of Rangeland and Watershed Management, Faculty of Natural Resources and Desert Studies, University of Yazd, Yazd, Iran.

2- Professor, Department of Rangeland and Watershed Management, Faculty of Natural Resources and Desert Studies, University of Yazd, Yazd, Iran. Department of Water Engineering, Faculty of Agriculture, Isfahan University of Technology, Isfahan, Iran.

3- Professor, Department of Rangeland and Watershed Management, Faculty of Natural Resources and Desert Studies, University of Yazd, Yazd, Iran.

4 - Associate Professor, Department of Rangeland and Watershed Management, Faculty of Agriculture, University of Ilam, Ilam, Iran.

* Corresponding Author: talebisf@yazd.ac.ir

Received: 2024/04/30

Accepted: 2024/06/25

the reliability of hydrological modeling. The findings of this study highlight the importance of simultaneously using multiple outlier detection techniques to improve statistical analyses and increase the accuracy of flood prediction in watersheds prone to extreme hydrological events.

Results:

Descriptive analysis of the hydrological data for the Godarkhush River indicated that for two key variables of annual maximum discharge and annual maximum precipitation, the skewness and kurtosis values were 3.87 and 17.53 for discharge, and 2.11 and 5.01 for precipitation, respectively. These values reflect a distribution with notable positive skewness and high kurtosis, commonly indicative of the presence of outliers in the dataset. Outliers can significantly affect statistical analyses and hydrological modeling; thus, their identification is crucial. To detect outliers, multiple methods were employed, including Box Plot, Z-Score method, and Cluster Analysis. The results consistently identified the discharge data for the year 2015, with a value of 2,160 m³/s, as an outlier across all methods. Subsequently, to determine the most suitable probability distribution for modeling annual maximum discharges, four common distributions (Log-Pearson Type III, Log-Normal, Weibull, and Gumbel Max) were evaluated using the EasyFit statistical software. According to the Anderson–Darling test, which is particularly sensitive to deviations in the tails of the distribution, the Log-Pearson Type III distribution yielded the lowest statistic (0.341), indicating the best fit. Similarly, based on the Chi-Squared test, this distribution also showed the best agreement with the observed data. Therefore, Log-Pearson Type III was selected as the most appropriate distribution for modeling annual peak discharge data in this study. To estimate design discharge values using the Log-Pearson Type III distribution, key distribution parameters (mean, standard deviation, and skewness) were calculated for the Takht-e-Khan hydrometric station under both conditions of with and without the identified outlier. The results showed that the inclusion of the outlier increased the mean annual discharge from 231.87 to 288.59 m³/s, highlighting the significant impact of extreme flood events on computed values. The skewness also increased by 1.77 units when the outlier was included. Moreover, kurtosis rose from 4.29 to 17.53 under the same condition, suggesting that including rare extreme events can considerably alter design flow estimations. Following this, design discharge values were calculated using the Log-Pearson Type III distribution for various return periods (5 to 1000 years), both with and without the outlier. The results revealed that for short return periods, the differences in design discharge between the two scenarios were minimal. However, for medium-term return periods (25 to 100 years), the disparity increased with the return period. For instance, at a 50-year return period, the design discharge was estimated at 1,481.5 m³/s with the outlier and 894.15 m³/s without it that indicated a 39.64% reduction when the outlier was excluded. Similarly, in long-term return periods (200 to 1000 years), the differences became even more pronounced. For the 1000-year return period, the design discharge was calculated at 7,152.4 m³/s with the outlier, and 2,532.9 m³/s without it, reflecting a 64.58% decrease that demonstrated the substantial influence of outlier events on long-term hydrological estimates.

Conclusion:

The findings of the study emphasize that managing outlier data and making the correct decision about including or excluding it from datasets, especially in flood analysis and design discharge estimation, can significantly affect the accuracy of predictions and hydrological designs. Based on the results, the flood discharge data for the year 2015 (2160 m³/s) was identified as an outlier. Due to its significant deviation from other data, this outlier can cause major changes in the design discharge estimates. Excluding this data from the dataset leads to a considerable reduction in the design discharge for long return periods, which could result in underestimating the risk of severe floods. Moreover, including the discharge data in 2015, as an exceptional data point, in the calculations improves the accuracy of predictions and the design of hydrological structures. These results highlight the importance of proper management of outlier data and the correct decision-making process regarding its inclusion or exclusion in order to avoid significant errors in predictions and designs.

Keywords: Design discharge, Uutlier data, Rare event, Return period, Statistical distribution.

Citation: Akbari, Z., Talebi, A., Ekhtesasi, M. R., Rostami, N., 2025. Analysis of the role of low-probability events on the accuracy of discharge estimation for different return periods. *Iranian Water Research Journal*. 58(3). pp.13-27. <https://dx.doi.org/10.22034/iwrj.2025.15193.2680>.



تحلیل نقش رخدادهای کم‌احتمال بر دقت برآورد دبی در دوره بازگشت‌های مختلف (مطالعه موردی: حوضه رودخانه گدارخوش ایلام)

زینب اکبری^۱، علی طالبی^{۲*}، محمدرضا اختصاصی^۳، نورالدین رستمی^۴

چکیده

سیلاب از بلایای طبیعی با تأثیرات گسترده بر زندگی انسان و محیط‌زیست است که تحلیل دقیق دبی‌های طراحی در دوره‌های بازگشت مختلف برای مدیریت مؤثر آن ضرورت دارد. رخدادهای کم‌احتمال یا داده‌های پرت، داده‌هایی هستند که به‌طور قابل‌توجهی از سایر اعضای آن جمعیت انحراف دارند و می‌توانند نتایج تحلیل‌های آماری را به مقدار زیادی تحت تأثیر قرار دهد. این مطالعه با هدف تحلیل دبی جریان و برآورد سیلاب‌های طراحی در حوضه آبخیز رودخانه گدارخوش، با استفاده از داده‌های بارش ایستگاه سینوپتیک ایلام و دبی ایستگاه هیدرومتری تخت‌خان، در دوره زمانی ۳۴ ساله (۱۳۹۹-۱۳۶۶) در شرایط وجود و عدم وجود داده پرت در مجموعه داده‌ها انجام گرفته است. برای شناسایی داده‌های پرت از روش‌های گرافیکی، آماری و روش خوشه‌بندی داده‌ها استفاده شد. در تحلیل دوره‌های بازگشت مختلف، بر اساس آزمون‌های برازش نکویی آندرسون-دارلینگ و کای‌اسکوئر، توزیع لوگ‌پیرسون نوع ۳ برای مدل‌سازی دبی انتخاب شد. نتایج نشان داد که دبی سال ۱۳۹۴ در تمام روش‌ها داده پرت است. مقایسه دبی طراحی برآورد شده با استفاده از توزیع آماری لوگ‌پیرسون نوع ۳ در شرایط با و بدون داده پرت نشان داد که در دوره‌های بازگشت کوتاه‌مدت اختلاف مقادیر برآورد شده کمتر و حدود ۰/۳۱-٪ است؛ در حالی که در دوره‌های بلندمدت اختلاف قابل‌توجهی به‌ویژه در دوره بازگشت ۱۰۰۰ ساله (۶۴/۵۸-٪) مشاهده شد. این یافته‌ها بر اهمیت مدیریت داده‌های پرت و لزوم استفاده از آن‌ها به عنوان داده نادر یا رخدادهای کم‌احتمال در تحلیل‌های هیدرولوژیکی و برآورد دبی طراحی تأکید دارد.

واژه‌های کلیدی: توزیع آماری، رخداد کم‌سابقه، داده پرت، دبی طراحی، دوره بازگشت

ارجاع: اکبری، ز.، طالبی، ع.، اختصاصی، م.، ر.، رستمی، ن.، ۱۴۰۴. تحلیل نقش رخدادهای کم‌احتمال بر دقت برآورد دبی در دوره بازگشت‌های مختلف. مجله پژوهش آب ایران، ۵۸(۳)، صص. ۲۷-۱۳. <https://dx.doi.org/10.22034/iwrj.2025.15193.2680>

^۱ - دانشجوی دکتری علوم و مهندسی آبخیز، گروه مرتع و آبخیزداری، دانشکده منابع طبیعی و کویرشناسی، دانشگاه یزد، یزد، ایران.

^۲ - استاد گروه مرتع و آبخیزداری، دانشکده منابع طبیعی و کویرشناسی، دانشگاه یزد، استاد گروه مهندسی آب، دانشکده کشاورزی، دانشگاه صنعتی اصفهان، اصفهان، ایران.

^۳ - استاد گروه مرتع و آبخیزداری، دانشکده منابع طبیعی و کویرشناسی، دانشگاه یزد، ایران.

^۴ - دانشیار گروه مرتع و آبخیزداری، دانشکده کشاورزی، دانشگاه ایلام، ایلام، ایران.

* نویسنده مسئول: talebisf@yazd.ac.ir

تاریخ پذیرش: ۱۴۰۳/۰۴/۰۵

تاریخ دریافت: ۱۴۰۳/۰۲/۱۱

مقدمه

سیلاب‌ها بلایای طبیعی هستند که تهدیدی جدی برای زندگی انسان‌ها و موجودات زنده به شمار می‌آیند. در سال‌های اخیر، سیلاب‌ها به‌طور فزاینده‌ای بسیاری از مردم در سراسر جهان را تحت تأثیر قرار داده و خسارات شدیدی نیز به بار آورده‌اند (Bhat et al., 2019; Doocy et al., 2013). برای محافظت از مردم و اموال در برابر سیلاب‌ها، لازم است که تدابیر سازه‌ای و غیرسازه‌ای در مناطقی که در معرض سیلاب قرار دارند، اجرا شود. درک میزان و فراوانی سیلاب‌ها در منطقه می‌تواند به برنامه‌ریزی برای کنترل سیلاب و حوادث ناشی از آن کمک کند. داده‌ها و اطلاعات دقیق از فراوانی و شدت وقوع سیلاب‌ها در یک منطقه برای طراحی سازه‌هایی مانند سدها، پل‌ها و جوی‌ها در مناطق آسیب‌پذیر ضروری است (Machado et al., 2015). ارزیابی ریسک سیلاب و طراحی تدابیر حفاظتی اغلب به یک برآورد قابل‌اعتماد از جریان‌های شدید با احتمال وقوع مشخص (مانند سیلاب با احتمال سالانه ۱٪) نیاز دارد (Okoli et al., 2018). جامعه مهندسی معمولاً این جریان‌های شدید برآورد شده را به‌عنوان «سیلاب‌های طراحی» می‌شناسد، زیرا بر برخی پارامترهای کلیدی (اندازه، ابعاد، هزینه و ایمنی) که برای هر زیرساخت مرتبط با آب بهینه‌سازی می‌شوند، تأثیر می‌گذارند (Rasekh et al., 2010).

دو روش اصلی که در حال حاضر برای برآورد سیلاب‌های طراحی استفاده می‌شوند عبارت‌اند از ۱- روش‌های آماری (معمولاً به‌عنوان تحلیل فراوانی سیلاب شناخته می‌شوند) که اغلب شامل برازش یک تابع توزیع احتمالی، مانند توزیع مقدار حدی تعمیم‌یافته، به یک سری از جریان‌های حداکثر سالانه اندازه‌گیری شده در یک ایستگاه هیدرومتری است. علاوه بر این روش کلاسیک مبتنی بر حداکثرهای سالانه، روش مقادیر حداکثر بیش از آستانه (PoT^1) نیز یکی از رویکردهای مهم و پرکاربرد در چارچوب نظریه مقادیر حدی (EVT^2) است. در این روش، به جای استفاده از تنها یک مقدار حداکثر در هر سال، از تمامی مقادیر جریان‌هایی که از یک آستانه بالا عبور کرده‌اند استفاده می‌شود و این مقادیر فراتر از آستانه با استفاده از توزیع پارتو تعمیم‌یافته (GPD^3) مدل‌سازی می‌گردند. روش PoT امکان بهره‌برداری از اطلاعات بیشتر نسبت به روش حداکثرهای سالانه را فراهم می‌سازد و به‌ویژه در تحلیل رخدادهای نادر و شدید، دقت بالاتری ارائه می‌دهد (Pipiras, 2020). ۲- روش‌های هیدرولوژیک که اغلب شامل برآورد

سیلاب‌های طراحی بر اساس استفاده از یک مدل ریاضی هستند که فرآیندهایی که در تبدیل بارش به رواناب در یک حوضه آبخیز مشخص نقش دارند، را توصیف می‌کند (Okoli et al., 2019). برای تحلیل فراوانی سیلاب، معمولاً از دبی‌های اوج استفاده می‌شود؛ اما در مواقعی که حوضه آبریز فاقد سابقه طولانی از داده‌های ایستگاه اندازه‌گیری دبی است، از داده‌های بارندگی نیز استفاده می‌شود. دوره مناسب برای تحلیل فراوانی سیلاب، یک دوره ۳۰ ساله از داده‌های حداکثر سالانه است (Khawairakpam, 2024). تحلیل فراوانی سیلاب را می‌توان با دو رویکرد مختلف انجام داد: تحلیل‌های محلی و تحلیل‌های منطقه‌ای. تحلیل فراوانی سیلاب در محل که نیاز به داده‌های ثبت‌شده طولانی‌مدت جریان دارد، رایج‌ترین رویکرد است. به‌طورمعمول، از تحلیل فراوانی سیلاب منطقه‌ای برای به دست آوردن برآوردهای قابل‌اعتماد از وقوع سیلاب و برای حوضه‌های آبخیز فاقد آمار استفاده می‌شود (Mangukiya and Sharma, 2024). با این حال، تحلیل فراوانی سیلاب محلی می‌تواند اطلاعات دقیق‌تری برای طراحی سازه‌های هیدرولیکی برای پیش‌بینی دقیق سیلاب‌های اوج ارائه دهد. جمع‌آوری تمام داده‌های هیدرولوژیکی به‌صورت کمی اغلب فرآیندی دشوار است. از سوی دیگر، انجام پیش‌بینی‌ها از طریق روش‌های آماری که از داده‌های اندازه‌گیری‌شده استفاده می‌کنند، مزایای بسیاری دارد (Hamzah et al., 2021). برای این منظور، دوره‌های بازگشت سیلاب را می‌توان با به‌کارگیری توابع توزیع احتمالات در مطالعات تحلیل فراوانی سیلاب محاسبه کرد. بر اساس داده‌های جریان ثبت‌شده، می‌توان مقادیر دبی را برای دوره‌های بازگشت مختلف به‌دست آورد. تحلیل داده‌هایی که طی سال‌های متمادی از مشاهدات به‌دست آمده‌اند، به دلیل دقت بالای نتایج، نقش بسزایی خواهد داشت (Turhan and Değerli, 2022). در تحلیل دبی با دوره‌های بازگشت مختلف، معمولاً از توزیع‌های آماری مانند توزیع گمبل، لوگ‌نرمال و پیرسون نوع ۳ برای مدل‌سازی داده‌های هیدرولوژیکی استفاده می‌شود. این توزیع‌ها به‌ویژه در شرایطی که داده‌های کم‌سابقه در دسترس هستند، برای برآورد دبی‌های حدی و وقوع سیلاب‌های نادر اهمیت ویژه‌ای دارند. داده‌های کم‌سابقه در هیدرولوژی به داده‌هایی اطلاق می‌شود که در طول دوره‌های زمانی طولانی‌مدت به‌طور نادر رخ می‌دهند و معمولاً حاوی اطلاعات با ارزشی برای تحلیل‌های سیلاب و دبی‌های حدی هستند. این داده‌ها نقش اساسی در مدل‌سازی سیلاب، برآورد

دبی با دوره‌های بازگشت مختلف و طراحی سازه‌های هیدرولوژیکی ایفا می‌کنند. در تحلیل‌های هیدرولوژیکی، دقت برآورد دبی‌ها با دوره‌های بازگشت متفاوت به شدت تحت تأثیر کیفیت و جامع بودن داده‌ها قرار دارد. به‌ویژه، داده‌های نادر و کم سابقه ممکن است بر انتخاب توزیع‌های آماری مناسب برای مدل‌سازی و پیش‌بینی تأثیرگذار باشند. رویدادهای تاریخی که پیش از سوابق سیستماتیک اتفاق افتاده‌اند، می‌توانند بینش‌هایی در مورد شدیدترین رویدادها فراهم کنند و امکان استنتاج بهتر این سیلاب‌های آسیب‌زا را فراهم آورند (Griffin et al., 2018). یکی از چالش‌های اصلی در استفاده از داده‌های کم‌سابقه، مدیریت داده‌های پرت^۱ است که ممکن است موجب تأثیر منفی بر برآورد دبی و تحلیل توزیع‌های آماری شود. روش‌های مختلف آماری به‌منظور شناسایی و حذف داده‌های پرت پیشنهاد شده‌اند (Rousseeuw and Leroy, 1987). در عین حال، به دلیل اهمیت بالای داده‌های نادر در طراحی‌های مهندسی، استفاده از این داده‌ها به‌طور مؤثر نیازمند تحلیل دقیق و انتخاب مدل‌های آماری مناسب است. مطالعات متعددی بر اهمیت استفاده از داده‌های تاریخی و مدل‌های آماری و هیدرولوژیکی برای برآورد دقیق سیلاب‌های طراحی تأکید کرده‌اند. نتایج مطالعه Saghaefian et al. (2013) نشان داد بارش‌های سیلابی تأثیر فراوانی بر نتایج شبیه‌سازی بارش-رواناب دارد و میزان این تأثیر با افزایش دوره بازگشت افزایش می‌یابد. Heydarpour et al. (2015) اثر سیلاب‌های استثنایی بر تحلیل فراوانی سیلاب در حوضه آبخیز سد گلستان را بررسی کردند. ابتدا سیلاب‌های استثنایی با استفاده از آزمون‌های پرت (دیکسون، گرایز و گرابز-بک) شناسایی شدند. سپس تحلیل‌های فراوانی سیلاب با سری کامل و با حذف سیلاب‌های استثنایی با استفاده از ۱۵ تابع توزیع احتمالاتی و سه روش برآورد (گشتاورها، بیشینه درست‌نمایی و گشتاورهای وزنی) انجام دادند. نتایج نشان داد توزیع لوگ‌پیرسون نوع ۳ در هر دو حالت به‌عنوان برترین توزیع انتخاب شده است. به همین ترتیب حذف سیلاب‌های استثنایی نوع توزیع را تغییر نمی‌دهد، اما تأثیر چشمگیری بر مقادیر برآوردی دارد. همچنین Mahmoodi and Mesbahi (2016) به تحلیل فراوانی سیلاب در رودخانه‌های سیوند و کر پرداختند و حداکثر دبی سیلاب و هیدروگراف‌های سیلاب را برای ۴ ایستگاه هیدرومتری با داده‌های ۳۵ تا ۴۶ ساله تعیین کردند. پس از انجام آزمون‌های

آماری، توزیع لوگ‌پیرسون نوع ۳ در ایستگاه‌های چمریز و بلاغی و لوگ‌نرمال ۳ پارامتری در ایستگاه‌های دشتبال و پل خان انتخاب و دبی سیلاب و هیدروگراف را برای دوره‌های بازگشت مختلف محاسبه نمودند.

(Strupczewski, 2014) با استفاده از داده‌های تاریخی و روش حداکثر درست‌نمایی، دقت تخمین چندک‌های بزرگ را بررسی کرده و نشان داد که گنجاندن طول مؤثر دوره تاریخی می‌تواند تخمین‌ها را بهبود بخشد. (Bacova Mitkova et al., 2021) نیز با استفاده از توزیع لوگ‌پیرسون نوع ۳ در رودخانه دانوب، نشان دادند که داده‌های تاریخی و منطقه‌ای سازی ضریب کشیدگی توزیع، تأثیر زیادی بر مقادیر طراحی جریان دارد. (Lefebvre, 2022) مدل فرآیند پواسن را برای تخمین دوره‌های بازگشت معرفی کرد که تطابق خوبی با داده‌های تاریخی داشت و امکان تحلیل اثرات تغییرات اقلیمی را فراهم کرد. (Ramasamy and Nagan, 2022) شش روش موقعیت‌نگاری (آداموسکی، هیزن، برد، چگودایف، توکی و برنارد) را همراه با مدل رگرسیون خطی لگاریتمی برای پیش‌بینی دبی سیلاب با دوره‌های بازگشت مختلف ارزیابی و از روش تحلیلی گمبل نیز به عنوان معیار مقایسه استفاده کردند. نتایج نشان داد روش هیزن با ضریب تعیین ۰/۹۲ نزدیک‌ترین پیش‌بینی را به روش گمبل ارائه می‌دهد. (Wu et al., 2023) تأثیر تغییرات اقلیمی را در چین بررسی کرده و دریافتند که اوج دبی و حجم سیلاب طراحی به‌طور قابل‌توجهی افزایش یافته است و بر اهمیت استفاده از داده‌های بارش سنگین و مطالعات منطقه‌ای تأکید کردند. (Khwaierakpam, 2024) مدل لوگ‌پیرسون نوع ۳ را به‌عنوان بهترین مدل برای تخمین سیلاب‌های اوج در دوره‌های بازگشت مختلف در رودخانه کوری هند معرفی کرد. هدف این تحقیق بررسی و شناسایی داده‌های پرت در مجموعه داده‌های حداکثر دبی سالانه ایستگاه هیدرومتری تخت‌خان، واقع در حوضه رودخانه گدارخوش ایلام است. در این راستا، تأثیر حذف یا ابقاء این داده‌ها به عنوان داده‌های پرت ارزیابی شده و نقش آن‌ها در قالب رخدادهای کم احتمال بر دقت برآورد دبی در دوره‌های بازگشت مختلف، با استفاده از توزیع‌های آماری، تحلیل و ارزیابی خواهد شد.

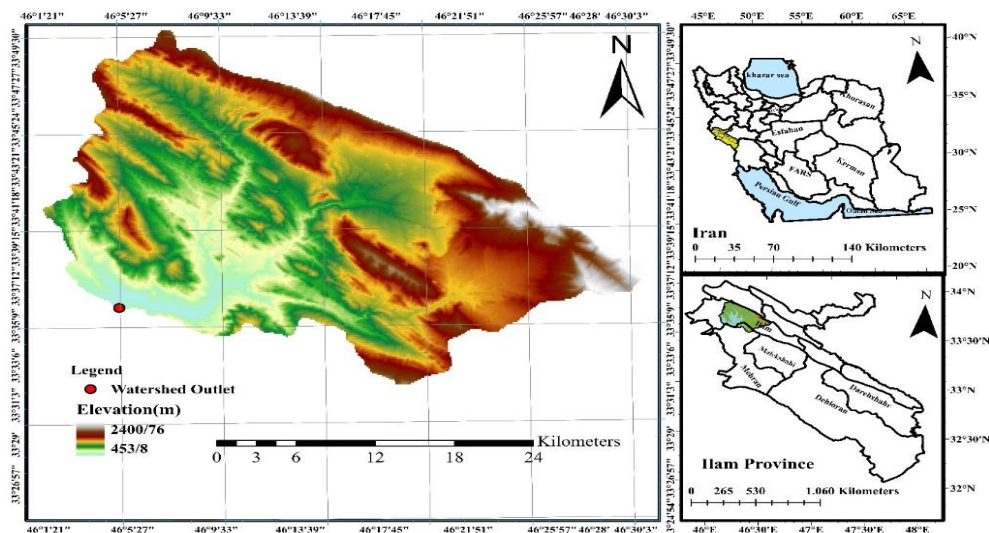
مواد و روش‌ها

منطقه مورد مطالعه

پیوستن این سرشاخه‌ها، جریان رودخانه به سمت جنوب غرب ادامه یافته و در حوالی روستای چم‌سارد با نام رودخانه گدارخوش شناخته می‌شود (Houshmand, 2005).

بیشتر حوضه آبریز رودخانه از زمین‌های تپه ماهوری، تیغ‌های فرسایشی و گیاهان پراکنده تشکیل شده است و تنها در نزدیکی مرز، قسمت‌هایی از دشت دارای شیب ملایمی می‌باشد. سرچشمه رودخانه از سنگ‌های آسماری متعلق به دوره میوسن و الیگوسن بوده که مسیر رودخانه موازی با ساختمان زمین است. در قسمت‌های پایین‌تر رودخانه، دره‌های باریک و تنگ سنگ‌های نرم‌تر را قطع می‌کنند. بهترین سازندهای منطقه آسماری، پابده، ایلام و گورپی هستند و سازندهای مهم در کیفیت آب آغاجاری، بختیاری، گچساران و رسوبات عصر حاضر هستند (Ramazani et al., 2014).

رودخانه گدارخوش از رودخانه‌های مرزی کشور در مرز ایران و عراق و در استان ایلام واقع شده است. این حوضه دارای مختصات جغرافیایی $31^{\circ}28'28''$ تا $33^{\circ}50'19''$ عرض شمالی می‌باشد. حوضه آبریز این رودخانه با مساحت 810 کیلومتر مربع بین سه حوضه آبریز کنجان چم، کنگیر و تلخاب واقع شده است. محیط حوضه آبخیز 176464 متر و ارتفاع متوسط آن نیز برابر با 1176 متر از سطح دریا می‌باشد (شکل ۱). این رودخانه از ارتفاعات کوه گچان، قلندر، رنو، شره زول، پارده و کوه الله خدا سرچشمه گرفته و پس از عبور از شهرهای چوار، ایلام، روستای بانویزه و چم آب در حدفاصل دو پاسگاه انجیره و نی خضر وارد خاک عراق می‌گردد. میزان آبدی رودخانه در طول سال و در فصول مختلف دچار نوسان است. این رودخانه از سه سرشاخه اصلی، شامل چوار، تنگ‌گراب و گلال‌رود تشکیل می‌شود. پس از



شکل ۱- موقعیت حوضه رودخانه گدارخوش در کشور و استان ایلام

داده‌های پرت یا رخ دادهای کم احتمال

داده‌های پرت داده‌هایی هستند که به‌طور قابل‌توجهی از سایر اعضای آن جمعیت انحراف دارند که نشان می‌دهد این عدد یا اعداد توسط روش یا سازوکار دیگری تولید شده است. (Rostampour, 2022; Ahmadi and Sarmadi, 2010) داده‌های پرت می‌تواند نتایج تحلیل‌های آماری را به مقدار زیادی تحت تأثیر قرار دهد، به‌طوری‌که حذف این عدد از مجموعه داده‌ها نتایج کاملاً متفاوتی به دست بیاورد. به همین دلیل شناسایی و تشخیص داده‌های پرت و بررسی اثر آن‌ها در یک تحلیل آماری از اهمیت خاصی برخوردار است. تنها یک داده پرت می‌تواند کل نتایج یک تحلیل آماری را بی‌اعتبار کند (Dadkhah and Samadi, 2018). در تحلیل‌های آماری

داده‌های مورد استفاده

در مطالعات هیدرولوژیکی و مدیریت حوضه‌های آبخیز، تحلیل دقیق رفتار جریان‌های سطحی و برآورد سیلاب‌های طراحی نیازمند استفاده از داده‌های جامع و دقیق است. این داده‌ها نقش اساسی در شبیه‌سازی فرآیندهای هیدرولوژیکی و پیش‌بینی دبی جریان با دوره‌های بازگشت مختلف دارند. در این پژوهش، مجموعه‌ای از داده‌های هواشناسی و هیدرومتری شامل آمار بارندگی برگرفته از ایستگاه سینوپتیک ایلام و آمار حداکثر دبی سالانه (ایستگاه هیدرومتری تخت‌خان) برای بازه زمانی ۳۴ سال (۱۳۶۶-۱۳۹۹)، به‌منظور تحلیل دبی جریان با دوره‌های بازگشت مختلف گردآوری شده است.

نمودار پراکندگی^۲: یک نمودار گرافیکی است که رابطه بین دو متغیر را نمایش می‌دهد. این نمودار شامل نقاطی است که هر نقطه نمایانگر یک جفت مقدار از دو متغیر است. یک نمودار پراکندگی یا نقطه‌ای ساده می‌تواند برای تعیین خطی بودن یک رابطه، تشخیص نقاط پرت و نمایش گرافیکی رابطه بین دو متغیر پیوسته استفاده شود.

نقاط رسم شده در نمودار پراکندگی نه تنها نقاط داده را گزارش می‌کند، بلکه الگوی پراکندگی داده‌ها را به صورت کلی بیان می‌کند و با یک نگاه می‌توان وضعیت داده‌ها را نسبت به هم بررسی و نقاطی که به طور قابل توجهی از بقیه داده‌ها فاصله دارند را پیدا کرد. در تحقیقات، داده‌های پرت می‌توانند به مشکلات یا فرصت‌های غیرمنتظره اشاره کنند. نمودار پراکندگی کمک می‌کند تا این نقاط غیرعادی شناسایی و دلایل احتمالی آن‌ها بررسی شود.

۲- روش‌های آماری

روش Z-Score: یکی از متداول‌ترین و ساده‌ترین روش‌ها برای شناسایی داده‌های پرت در مجموعه‌های داده‌ای است. این روش بر اساس استاندارد کردن داده‌ها عمل می‌کند تا هر داده در مقیاس واحد انحراف معیار قرار گیرد. در این روش، Z-Score نشان‌دهنده میزان انحرافات معیار یک داده از میانگین داده‌ها است. Z-Score برای یک داده x به صورت رابطه ۱ محاسبه می‌شود.

$$Z = \frac{x_i - \mu}{\sigma} \quad (1)$$

که در آن، x_i داده مورد نظر، μ میانگین داده‌ها و σ انحراف معیار داده‌ها است.

برای هر داده در مجموعه، Z-Score آن محاسبه می‌شود. این مقدار نشان‌دهنده فاصله داده از میانگین در واحد انحراف معیار است. برای شناسایی داده‌های پرت، معمولاً از یک آستانه برای Z-Score استفاده می‌شود. به طور معمول، اگر Z-Score یک داده بزرگ‌تر از ۳ یا کوچک‌تر از -۳ باشد، آن داده به عنوان داده پرت شناسایی می‌شود (Peruri Venkata et al., 2019).

این به این معنی است که داده‌هایی که Z-Score آن‌ها بیشتر از ۳ انحراف معیار از میانگین داده‌ها فاصله دارند، به احتمال زیاد داده‌های پرت هستند.

۳- روش‌های مبتنی بر مدل

تحلیل خوشه‌ای^۳: تحلیل خوشه‌ای یک تحلیل اکتشافی است که سعی می‌کند ساختارهای درون داده‌ها را شناسایی کند.

تشخیص داده‌های پرت به دو دلیل از اهمیت زیادی برخوردار است ۱- وجود داده‌های پرت در یک مجموعه داده اثر نامطلوبی بر استنباط آماری از آن مجموعه داده می‌گذارد و ۲- در برخی موارد داده‌های پرت اطلاعات مفیدی از مدل داده‌ها به تحلیل‌گر ارائه می‌دهد؛ از این رو شناسایی آن برای تحلیل‌گر حائز اهمیت است (Dehghan and Faridrohani, 2022; Rostampour, 2022). در زمینه برآورد دبی با دوره‌های بازگشت مختلف، وجود و یا عدم وجود داده‌های پرت می‌تواند تأثیر قابل توجهی بر پارامترهای آماری توزیع‌های احتمالی مانند میانگین، انحراف معیار و پارامترهای حدی داشته باشد (Kite, 1997). این موضوع می‌تواند به پیش‌بینی‌های نادرست از دبی‌های حدی و طراحی غیر ایمن سازه‌های هیدرولیکی، مانند سدها و کانال‌ها منجر شود (Nathan and McMahon, 1990). به همین دلیل، شناسایی و مدیریت داده‌های پرت، گامی اساسی در بهبود کیفیت داده‌ها و افزایش دقت تحلیل‌های هیدرولوژیکی است.

روش‌های شناسایی داده‌های پرت در هیدرولوژی

روش‌های مختلفی برای شناسایی و حذف داده‌های پرت در تحلیل داده‌های هیدرولوژیکی پیشنهاد شده است. انتخاب روش مناسب به ماهیت داده‌ها، توزیع آماری آن‌ها و اهداف تحلیل بستگی دارد. در این پژوهش، از روش‌های گرافیکی (تحلیل نمودارهای جعبه‌ای و پراکندگی)، روش‌های آماری (روش Z-Score و...) و روش‌های مبتنی بر مدل (تحلیل خوشه‌ای و...) استفاده شده است.

۱- روش‌های گرافیکی

نمودار جعبه‌ای^۱: در آمار توصیفی نمودار جعبه‌ای نموداری است که برای توصیف پراکندگی داده به کار می‌رود. در این نمودار از جعبه‌ای برای نمایش فاصله بین چارک اول و سوم استفاده می‌شود و خطی در داخل جعبه، میانه (چارک دوم) را مشخص می‌کند. خارج از جعبه مقادیر حداقل و حداکثر داده‌ها را نیز مشخص می‌نمایند. برای شناسایی داده‌های پرت، حدود پایین و بالای آستانه بر اساس روش چارک‌ها تعیین می‌شود. بدین منظور، حد پایین با کسر ۱/۵ برابر دامنه میان‌چارکی از چارک اول و حد بالا با افزودن ۱/۵ برابر دامنه میان‌چارکی به چارک سوم محاسبه می‌گردد. داده‌هایی که خارج از این بازه قرار گیرند، به عنوان داده‌های پرت شناسایی می‌شوند.

^۳- Cluster Analysis

1 -Box Plot
2 -Scatter Plot

$$f(x|\tau, \alpha, \beta) = \frac{\left(\frac{x-\tau}{\beta}\right)^{\alpha-1} \exp\left(-\frac{x-\tau}{\beta}\right)}{|\beta|\Gamma(\alpha)}$$

$$\frac{x-\tau}{\beta} \geq 0 \quad (3)$$

که در آن، τ پارامتر مکان، α پارامتر شکل، β پارامتر مقیاس و $\Gamma(\alpha)$ تابع گاما است که در رابطه ۴ ارائه شده است.

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} \exp^{-t} dt \quad (4)$$

روش گشتاور از لگاریتم متغیرها برای برآورد پارامترهای توزیع استفاده می‌کند $\log X = \mu + K\sigma$ که در آن X یک متغیر تصادفی است، μ میانگین، σ انحراف معیار و K عاملی وابسته به ضریب چولگی در یک احتمال تجاوز مشخص شده است.

شرایط سری Q-max

توزیع با محاسبه لگاریتم‌های پایه ۱۰ دبی (Q) در یک احتمال تجاوز مشخص (P)، برازش داده می‌شود که با استفاده از معادله محاسبه می‌گردد:

$$\log Q = \bar{X} + K\sigma \quad (5)$$

که در آن $\bar{X}$ ، میانگین، σ انحراف معیار و K عاملی از ضریب چولگی در احتمال تجاوز انتخاب شده است. فرمول‌های این پارامترها در روابط ۶ و ۷ ارائه شده‌اند.

$$\bar{X} = \frac{1}{n} \sum X_i \quad (6)$$

$$\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (7)$$

نتایج و بحث

در تحلیل‌های هیدرولوژیک و بارش، درک و ارزیابی ویژگی‌های توزیع داده‌ها از اهمیت ویژه‌ای برخوردار است زیرا ویژگی‌های آماری مانند میانگین، انحراف معیار، چولگی و کشیدگی می‌توانند اطلاعات مفیدی در مورد رفتار کلی داده‌ها و تطابق آن‌ها با مدل‌های نظری و پیش‌بینی‌های موردنظر ارائه دهند. جدول ۱ نتایج توصیفی دو متغیر مهم حداکثر دبی و حداکثر بارش سالانه داده‌ها را نشان می‌دهد. مقدار کشیدگی دو متغیر (به ترتیب ۱۷/۵۳ و ۵/۰۱) و چولگی (به ترتیب ۳/۸ و ۲/۱۱) نشان دهنده یک توزیع با چولگی مثبت شدید و کشیدگی بالا است که به طور معمول نشان‌دهنده حضور داده‌های پرت است که می‌توانند تأثیر قابل‌توجهی در تحلیل‌های آماری و مدل‌سازی‌های هیدرولوژیکی داشته باشند.

جدول ۱- آمار توصیفی متغیرهای حداکثر دبی سالانه ایستگاه هیدرومتری تخت‌خان و حداکثر بارش سالانه ایستگاه سینوپتیک ایلام در دوره زمانی ۲۴ ساله (۱۳۹۹-۱۳۶۶)

تجزیه و تحلیل خوشه‌ای را تجزیه و تحلیل تقسیم‌بندی یا تجزیه و تحلیل طبقه‌بندی نیز می‌نامند. به‌طور خاص، این روش سعی می‌کند گروه‌های همگن از داده‌ها را در صورتی که گروه‌بندی قبلاً شناخته‌شده نبود، شناسایی کند. چون تجزیه و تحلیل خوشه‌ای اکتشافی است، هیچ تمایزی بین متغیرهای وابسته و مستقل قائل نمی‌شود. روش‌های مبتنی بر خوشه‌بندی، خوشه‌های کوچک را به‌عنوان داده‌های پرت در نظر می‌گیرند. منظور از خوشه‌های کوچک، خوشه‌هایی هستند که تعداد داده‌های آن به‌میزان قابل‌توجهی کمتر از سایر خوشه‌ها باشد. روش فوق، بدون ناظر می‌باشد و می‌توان پس از خوشه‌بندی نقاط جدید را وارد و پرت بودن آن‌ها را بررسی کرد (Kiani and Montazeri, 2015).

توزیع آماری لوگ‌پیرسون نوع ۳^۱

توزیع لوگ‌پیرسون نوع ۳ یکی از بهترین گزینه‌ها برای تحلیل داده‌های حدی در هیدرولوژی است. با توجه به چولگی مثبت داده‌ها و وجود رویدادهای حدی مانند سیلاب‌های شدید، این توزیع ابزار قدرتمندی برای برآورد دبی طراحی و مدیریت ریسک سیلاب به شمار می‌رود. اغلب برای کاهش چولگی از داده‌ها لگاریتم گرفته می‌شود و سپس میانگین، انحراف معیار و ضریب چولگی، بر اساس لگاریتم داده‌ها به دست می‌آید و از توزیع پیرسون نوع ۳ استفاده می‌شود که در این صورت آن را توزیع لوگ‌پیرسون نوع ۳ می‌نامند (Mahdavi, 2005). این توزیع به‌طور گسترده برای تحلیل فراوانی سیلاب و برآورد دبی طراحی در دوره‌های بازگشت مختلف (مثلاً ۵۰، ۱۰۰ و ۵۰۰ ساله) استفاده می‌شود. دوره بازگشت به‌صورت معکوس احتمال عدم تجاوز رویداد (رابطه ۲) تعریف می‌شود، یعنی:

$$\frac{1}{P} = T \quad (2)$$

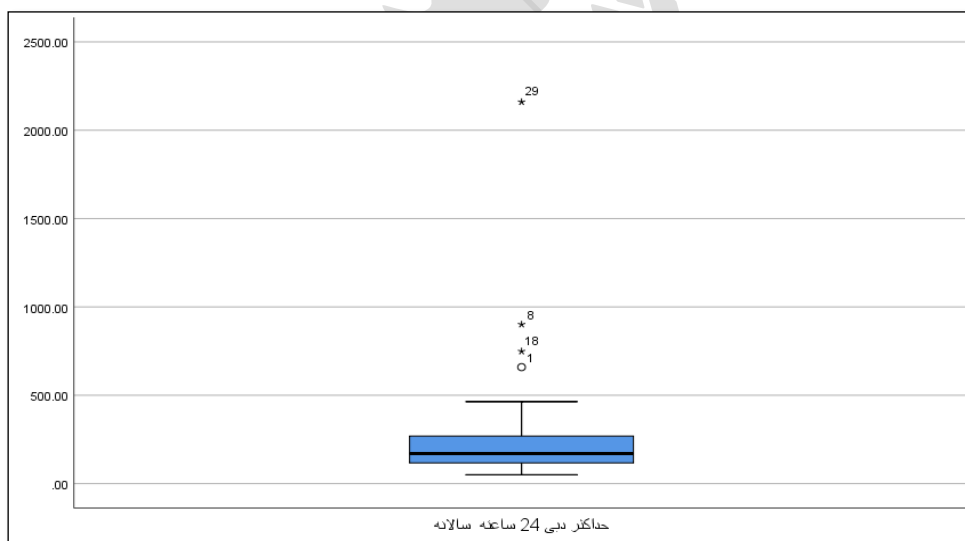
که در آن، T دوره بازگشت و P ، احتمال وقوع می‌باشد.

توزیع پیرسون نوع ۳ یک توزیع گاما با سه پارامتر است که یک تبدیل لگاریتمی از متغیر دارد. این توزیع به‌طور گسترده‌ای برای تحلیل سیلاب‌ها استفاده می‌شود، زیرا داده‌ها معمولاً به‌خوبی با سری‌های حداکثر سالانه جریان تخمینی تطابق دارند. تابع چگالی احتمال توزیع پیرسون نوع ۳ به‌صورت رابطه ۳ تعریف می‌شود:

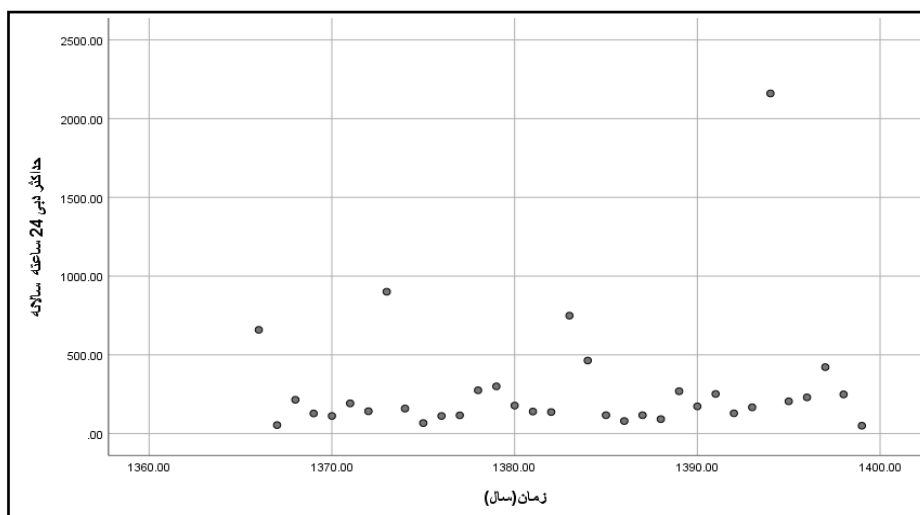
معیار	حداکثر بارش سالانه	حداکثر دبی سالانه
میانگین	۶۷/۹۱ (mm)	۲۸۸/۵۹ (m^3/s)
انحراف معیار	۳۶/۴۰	۳۸۴/۲۷
چولگی	۲/۱۱	۳/۸۸
کشیدگی	۵/۰۱	۱۷/۵۳
کمینه	۲۶ (mm)	۵۰ (m^3/s)
بیشینه	۱۸۸ (mm)	۲۱۶۰ (m^3/s)

در مرحله بعد، از روش محاسبه Z-Score به عنوان یک روش آماری، با بررسی انحراف داده‌ها از میانگین برای شناسایی داده‌های پرت استفاده گردید. به این صورت که اگر مقدار Z بزرگ‌تر از ۳ یا کوچک‌تر از -۳ باشد، داده پرت محسوب می‌شود. بر این مبنا و بر اساس مقادیر Z-Score محاسبه شده مطابق جدول ۲ دبی‌های ثبت شده مربوط به سال‌های ۱۳۸۳، ۱۳۷۳ و ۱۳۹۴ در شمار داده‌های پرت قرار گرفته‌اند. در گام بعدی، از روش تحلیل خوشه‌ای به عنوان یکی از روش‌های مبتنی بر مدل برای شناسایی داده‌های پرت استفاده شد. در این روش داده‌هایی که جدا از سایر خوشه‌ها قرار گیرند، جزء داده‌های پرت شناسایی می‌شوند. همانطور که در شکل ۴ نیز مشخص است، شاخه ۲۹ که مربوط به دبی سال ۱۳۹۴ می‌باشد، به عنوان داده پرت در یک شاخه جدا شده است.

برای شناسایی داده‌های پرت، از چند روش استفاده شد. ابتدا نسبت به شناسایی داده‌های پرت از طریق نمایش چارک‌ها و حذف‌ها در قالب ترسیم نمودار جعبه‌ای اقدام گردید. همان‌طور که در شکل ۲ نشان داده شده است، نقاط خارج از محدوده (دایره یا ستاره) نشان‌دهنده داده‌های پرت هستند که در این بین داده شماره ۲۹، ۸، ۱۸ و ۱ بیشترین فاصله را داشته که به ترتیب مربوط به دبی سال ۱۳۷۳، ۱۳۹۴، ۱۳۸۳ و ۱۳۶۶ می‌باشند. به‌طور واضح‌تر داده دبی مربوط به سال ۱۳۹۴ مشخصاً دورافتاده است. به همین ترتیب، در شکل ۳ که نمودار پراکندگی داده‌های حداکثر دبی سالانه را نشان می‌دهد دبی مربوط به سال ۱۳۹۴ با فاصله زیادی نسبت به داده‌های دیگر قرار گرفته است و در شمار داده‌های پرت قرار می‌گیرد.



شکل ۲- نمودار جعبه‌ای داده‌های حداکثر دبی سالانه ایستگاه هیدرومتری تخت‌خان در دوره زمانی ۳۴ ساله (۱۳۶۶-۱۳۹۹)



شکل ۳- نمودار پراکندگی داده‌های حداکثر دبی سالانه ایستگاه هیدرومتری تخت‌خان در دوره زمانی ۳۴ ساله (۱۳۶۶-۱۳۹۹)

جدول ۲- میزان Z-Score داده‌های حداکثر دبی سالانه ایستگاه هیدرومتری تخت‌خان در دوره زمانی ۳۴ ساله (۱۳۶۶-۱۳۹۹)

حداکثر دبی سالانه (m ³ /s)	Zscore حداکثر دبی سالانه	زمان (سال)	حداکثر دبی سالانه (m ³ /s)	Zscore حداکثر دبی سالانه	زمان (سال)
۶۵۹	۰/۹۶	۱۳۸۳	۷۴۹	*۱۱۹/۸	۱۳۶۶
۵۴	-۰/۶۱	۱۳۸۴	۴۶۴	۰/۴۵	۱۳۶۷
۲۱۵	-۰/۱۹	۱۳۸۵	۱۱۷	۰/۴۴	۱۳۶۸
۱۲۸	-۰/۴۱	۱۳۸۶	۸۰	-۰/۵۴	۱۳۶۹
۱۱۲	-۰/۴۶	۱۳۸۷	۱۱۷	-۰/۴۴	۱۳۷۰
۱۹۲	-۰/۲۵	۱۳۸۸	۹۲	-۰/۵۱	۱۳۷۱
۱۴۲	-۰/۳۸	۱۳۸۹	۲۶۹	-۰/۰۵	۱۳۷۲
۹۰۱	*۱۵۹/۳۶	۱۳۹۰	۱۷۳	-۰/۳	۱۳۷۳
۱۵۹	-۰/۳۳	۱۳۹۱	۲۵۲	-۰/۰۹	۱۳۷۴
۶۷	-۰/۵۷	۱۳۹۲	۱۲۹	-۰/۴۱	۱۳۷۵
۱۱۲	-۰/۴۶	۱۳۹۳	۱۶۷	-۰/۳۱	۱۳۷۶
۱۱۶	-۰/۴۴	۱۳۹۴	۲۱۶	*۴۸/۷	۱۳۷۷
۲۷۵	-۰/۳۵	۱۳۹۵	۲۰۵	-۰/۲۱	۱۳۷۸
۳۰۰	۰/۰۳	۱۳۹۶	۲۳۰	-۰/۱۵	۱۳۷۹
۱۷۸	-۰/۲۸	۱۳۹۷	۴۲۲	۰/۳۴	۱۳۸۰
۱۴۰	-۰/۳۸	۱۳۹۸	۲۴۹	-۰/۱	۱۳۸۱
۱۳۷	۰/۳۹	۱۳۹۹	۵۰	-۰/۶۲	۱۳۸۲

* داده‌های پرت شناسایی شده

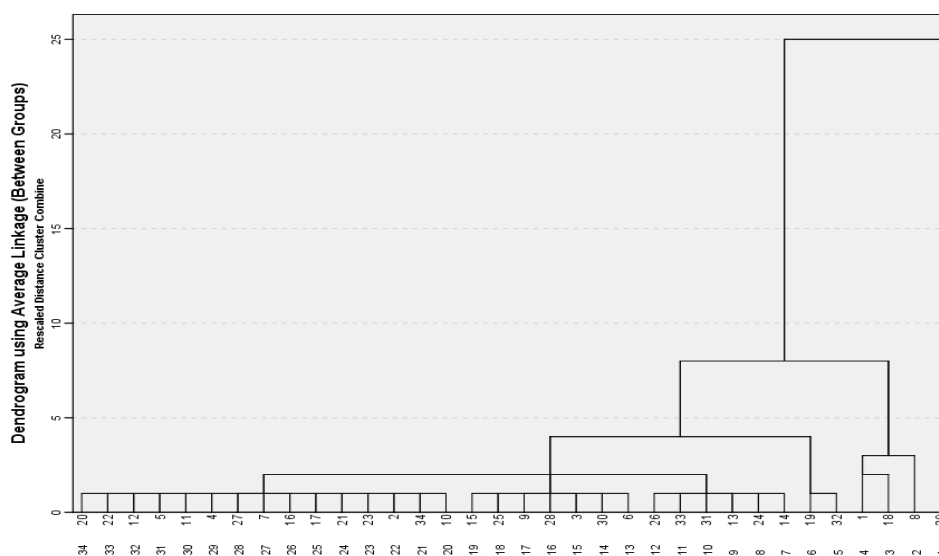
و این در حالی است که سوابق و اسناد موجود وقوع این سیلاب را اثبات کرده است. لذا برای دقت بیشتر بر موضوع، ضمن ترسیم سری زمانی داده‌های بارش و دبی (شکل ۵) نسبت به تعیین همبستگی بین داده‌های حداکثر بارش و حداکثر دبی سالانه بر اساس آزمون همبستگی اسپیرمن^۱ اقدام شد.

به‌طور کلی و بر اساس مقایسه روش‌های مختلف شناسایی داده‌های پرت که نتایج آن در قالب جدول ۳ ارائه شده است، مشخص گردید که دبی مربوط به سال ۱۳۹۴ با مقدار ۲۱۶۰ مترمکعب در ثانیه به صورت مشترک در همه روش‌ها به‌عنوان داده پرت شناسایی شده است و باید از بین داده‌ها حذف گردد

¹ Spearman Correlation

چرا که سیلاب طراحی باید بر مبنای داده‌های واقعی به وقوع پیوسته صورت گیرد تا طراحی سازه‌ها با شکست مواجه نشود. بر این اساس برای شناخت تأثیر این داده‌ها در برآورد مقدار دبی طراحی در دوره بازگشت‌های مختلف، اقدام به برآورد دبی با دوره بازگشت‌های مختلف با استفاده از توزیع‌های آماری لوگ‌پیرسون نوع ۳ در دو حالت وجود و یا عدم وجود داده پرت در مجموعه داده‌ها شد.

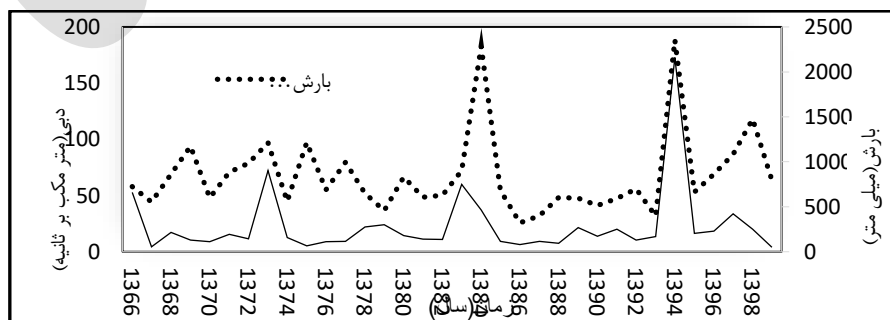
همان‌طور که در جدول ۴ مشاهده می‌شود مقدار T بزرگ‌تر از صفر و دارای همبستگی مثبت و مقدار p -value کمتر از ۰/۰۵ می‌باشد؛ بنابراین همبستگی معنی‌دار است. بررسی داده‌های حداکثر بارش سالانه نیز بیانگر بالاترین مقدار بارش در طول دوره آماری مورد بررسی می‌باشد. بنابراین تنها با استفاده از آزمون شناسایی داده‌های پرت نمی‌توان اقدام به حذف داده‌ها نمود و باید به‌عنوان داده نادر و کم‌احتمال با آن برخورد کرد



شکل ۴- نمودار خوشه‌ای داده‌های حداکثر دبی سالانه ایستگاه هیدرومتری تخت‌خان در دوره زمانی ۳۴ ساله (۱۳۹۹-۱۳۶۶)

جدول ۳- نتایج شناسایی داده‌های پرت جریان با استفاده از روش‌های گرافیکی، آماری و مبتنی بر مدل

داده‌های پرت شناسایی شده	روش آزمون							
	روش‌های گرافیکی				روش‌های آماری		روش‌های مبتنی بر مدل	
	نمودار جعبه‌ای		نمودار پراکندگی		Z-Score روش		تحلیل خوشه‌ای	
	سال	دبی (m^3/s)	سال	دبی (m^3/s)	سال	دبی (m^3/s)	سال	دبی (m^3/s)
	۱۳۶۶	۶۵۹	۱۳۹۴	۲۱۶۰	۱۳۷۳	۹۰۱	۱۳۹۴	۲۱۶۰
	۱۳۷۳	۹۰۱			۱۳۸۳	۷۴۹		
	۱۳۸۳	۷۴۹			۱۳۹۴	۲۱۶۰		
	۱۳۹۴	۲۱۶۰						



شکل ۵- سری زمانی داده‌های حداکثر دبی سالانه ایستگاه هیدرومتری تخت‌خان و حداکثر بارش سالانه ایستگاه سینوپتیک ایلام در دوره زمانی ۳۴ ساله (۱۳۹۹-۱۳۶۶)

جدول ۴- نتایج آزمون همبستگی داده‌های حداکثر دبی سالانه ایستگاه هیدرومتری تخت‌خان و حداکثر بارش سالانه ایستگاه سینوپتیک ایلام در دوره زمانی ۳۴ ساله (۱۳۶۶-۱۳۹۹)

آزمون همبستگی اسپیرمن- حداکثر دبی و بارش سالانه	
ضریب همبستگی r	۰/۳۴
p-value	۰/۰۴
خطای استاندارد	۰/۱۶
حد پایین سطح	-۰/۰۱۳
اطمینان	۰/۶۴
حد بالای سطح اطمینان	

است. همان‌طور که مشاهده می‌شود میانگین دبی سالانه با وجود داده پرت از مقدار $231/87$ به $288/59$ افزایش یافته، که نشان‌دهنده اثر قابل توجه وقایع سیلابی بر مقادیر محاسباتی است. ضریب چولگی نیز در شرایطی که داده پرت وارد محاسبات گردیده، به میزان $1/77$ بیشتر شده است. این نتایج مطابق با نتایج مطالعه Bačová Mitková et al (2021) می‌باشد که نشان دادند مقدار ضریب چولگی با اضافه کردن داده‌های پرت افزایش می‌یابد. افزایش انحراف معیار در شرایط وجود داده‌های پرت نشان‌دهنده پراکندگی بیشتر داده‌ها حول میانگین است که احتمالاً ناشی از وجود مقادیر حدی (سیلاب‌های بزرگ) در مجموعه داده‌هاست. افزایش مقدار کشیدگی نیز از $4/29$ به $17/53$ در شرایطی که داده پرت لحاظ شده است نشان‌دهنده این است که گنجاندن رخدادهای کم احتمال می‌تواند دبی‌های طراحی را تغییر دهد.

جدول ۶- میانگین، انحراف معیار و ضریب چولگی ایستگاه هیدرومتری تخت‌خان در شرایط وجود و حذف داده پرت

پارامتر آماری	حداکثر دبی سالانه در شرایط حذف داده	حداکثر دبی سالانه در شرایط وجود داده
میانگین	$231/87$	$288/59$
انحراف معیار	$198/79$	$384/27$
چولگی	$2/11$	$3/88$
کشیدگی	$4/29$	$17/53$
تعداد داده	۳۳	۳۴

در ادامه بر اساس تابع توزیع احتمال لوگ‌پیرسون نوع ۳ اقدام به محاسبه دبی طراحی با دوره بازگشت‌های مختلف (۵ تا ۷ و شکل ۶، مقایسه نتایج ارائه شده است. این مقایسه تأثیر داده پرت در برآورد دبی طراحی و دقت محاسبات هیدرولوژیکی

برآورد دبی با دوره بازگشت‌های مختلف بر اساس بهترین توزیع آماری

ابتدا به منظور انتخاب مناسب‌ترین توزیع آماری برای مدل‌سازی دبی‌های حداکثر سالانه، از آزمون‌های نکویی برازش آندرسون-دارلینگ و کای‌دو استفاده شد. در این مطالعه، چهار توزیع آماری رایج شامل لوگ‌پیرسون نوع ۳، لوگ‌نرمال، ویبول و گمبل بیشینه در محیط نرم افزار آماری Easyfit ارزیابی شدند. مقایسه این چهار توزیع براساس نتایج دو آزمون نکویی برازش مذکور در جدول ۵ ارائه شده است. همان‌طور که مشاهده می‌شود بر اساس آزمون آندرسون-دارلینگ، که حساسیت بیشتری به انحراف در انتهای توزیع دارد، توزیع لوگ‌پیرسون نوع ۳ در مقایسه با سایر توزیع‌ها با آماره $0/341$ کمترین مقدار را داشته است و بر حسب آزمون کای‌دو نیز برازش مناسب‌تری با داده‌های مشاهداتی دارد. بنابراین، توزیع لوگ‌پیرسون نوع ۳ در این پژوهش به عنوان مناسب‌ترین تابع برازش بر داده‌های حداکثر دبی سالانه انتخاب شد.

جدول ۵- نتایج آزمون نکویی برازش

روش آزمون		آندرسون-دارلینگ		کای اسکوئر	
توزیع آماری	آماره	رتبه	رتبه	آماره	رتبه
لوگ‌پیرسون نوع ۳	$0/341$	۱	۱	$0/62016$	۱
لوگ‌نرمال	$0/66376$	۲	۲	$6/310$	۲
ویبول	$2/307$	۳	۳	$8/6839$	۴
گمبل بیشینه	$3/98$	۴	۴	$7/3423$	۳

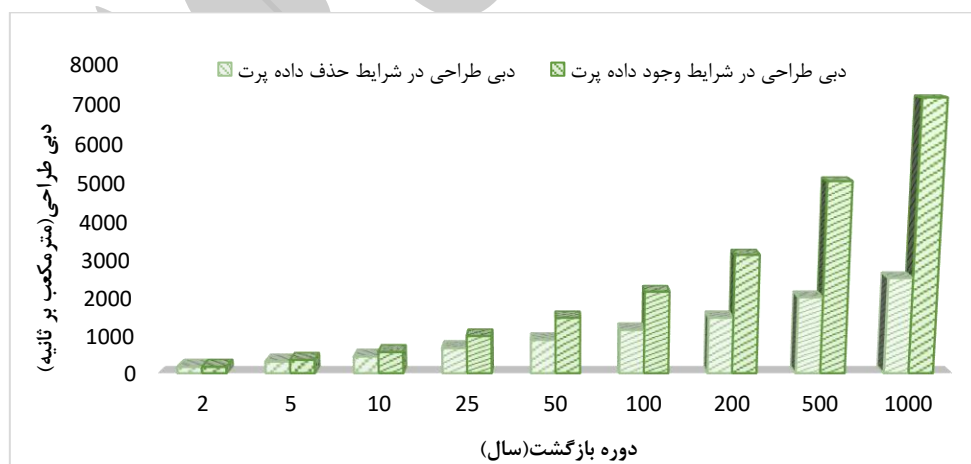
در ادامه پژوهش برای تخمین مقادیر دبی طراحی با استفاده از توزیع احتمالی لوگ‌پیرسون نوع ۳ برای شرایطی که داده پرت حذف و یا در مجموعه داده‌ها گنجانده شود، در اولین گام اقدام به محاسبه پارامترهای تابع توزیع شامل میانگین، انحراف معیار و ضریب چولگی برای ایستگاه تخت‌خان در شرایط حذف و حفظ داده پرت گردید که نتایج در جدول ۶ نشان داده‌شده

al در سال ۲۰۲۱ نیز نشان داد که در نظر گرفتن سیلاب‌های کم احتمال موجب تغییر در انحنای منحنی‌های توزیع لوگ‌پیرسون نوع ۳ و تغییر در دبی طراحی شده است. موافق با نتایج این پژوهش، Wu et al. (2023) نیز به این نتیجه رسیدند که با ورود داده‌های سیلاب‌های بزرگ مربوط به سال‌های ۲۰۱۰ و ۲۰۲۱، اوج دبی طراحی برای فراوانی‌های مختلف بین ۳/۲٪ تا ۵۰/۱٪ و حجم سیلاب ۲۴ ساعته طراحی بین ۹/۴٪ تا ۴۲/۲٪ افزایش یافته است. یافته‌های این پژوهشگران نشان داد که رخدادهای کم احتمال بر سیلاب طراحی در ایستگاه لیچینگ‌دیان تأثیر داشته‌اند؛ بنابراین، تحلیل و محاسبه سیلاب طراحی باید بر اساس استفاده کامل از داده‌های بارش سنگین و سیلاب‌های کم احتمال، همراه با داده‌های حاصل از مطالعات منطقه‌ای انجام شود تا دقت بیشتری در طراحی و مدیریت سیلاب حاصل شود.

را نشان می‌دهد، به این صورت که در دوره بازگشت‌های کوتاه مدت (۵ تا ۱۵ سال) اختلاف بین دبی طراحی در دو حالت کمتر است. این موضوع نشان‌دهنده تأثیر کمتر داده پرت در بازه‌های زمانی کوتاه است. در دوره بازگشت‌های میان مدت (۲۵ تا ۱۰۰ سال) با افزایش دوره بازگشت، اختلاف دبی طراحی در دو حالت بیشتر می‌شود. به طوری که در دوره بازگشت ۵۰ ساله، دبی طراحی با حضور داده پرت ۱۴۸۱/۵ مکعب بر ثانیه و بدون داده پرت ۸۹۴/۱۵ متر مکعب بر ثانیه است که به میزان ۳۹/۶۴ درصد کمتر برآورد شده است. به همین صورت در دوره بازگشت‌های بلندمدت (۲۰۰ تا ۱۰۰۰ سال) اختلاف‌ها به شدت افزایش می‌یابد، به گونه‌ای که در دوره بازگشت ۱۰۰۰ ساله، دبی طراحی با وجود داده پرت، ۷۱۵۲/۴ متر مکعب بر ثانیه و با حذف داده پرت ۲۵۳۲/۹ متر مکعب بر ثانیه محاسبه شده است که از نظر آماری ۶۴/۵۸٪ اختلاف داشته و این اختلاف بزرگ نشان‌دهنده تأثیر بسیار زیاد داده پرت در دوره بازگشت‌های بلندمدت است. تحقیق. Bačová Mitková et

جدول ۷- مقادیر دبی طراحی برای دوره بازگشت‌های ۵ تا ۱۰۰۰ سال با و بدون وجود داده پرت، بر اساس توزیع لوگ-پیرسون نوع ۳ در ایستگاه

هیدرومتری تخت‌خان									
دوره بازگشت (سال)	۵	۱۰	۱۵	۲۵	۵۰	۱۰۰	۲۰۰	۵۰۰	۱۰۰۰
دبی طراحی (وجود داده پرت)	۱۶۹/۷۹	۳۵۶/۸۹	۵۶۹/۰۷	۹۹۶/۱۴	۱۴۸۱/۵	۲۱۶۹	۳۱۳۷/۸	۵۰۴۰/۳	۷۱۵۲/۴
دبی طراحی (حذف داده پرت)	۱۶۹/۲۵	۳۱۴/۷۱	۴۵۰/۱۶	۶۷۶/۹۵	۸۹۴/۱۵	۱۱۶۰	۱۴۸۴/۵	۲۰۲۳/۶	۲۵۳۲/۹
تغییر دبی برآورد شده (%)	-۰/۳۱	-۱۱/۸۱	-۲۰/۸۹	-۳۲/۰۴	-۳۹/۶۴	-۴۶/۵۱	-۵۲/۶۸	-۵۹/۸۵	-۶۴/۵۸



شکل ۶- مقایسه مقادیر دبی طراحی در دوره‌های بازگشت مختلف در شرایط حضور و عدم حضور داده‌های پرت در ایستگاه هیدرومتری تخت‌خان

نتیجه‌گیری

داد، داده دبی سیلابی سال ۱۳۹۴ با مقدار ۲۱۶۰ مترمکعب در ثانیه، بالاترین مقدار دبی در میان سری داده‌های حداکثر سالانه بوده و در تمامی روش‌های شناسایی مورد استفاده به عنوان داده پرت شناخته شد. این داده به دلیل فاصله زیادی که از سایر

در تحلیل‌های هیدرولوژیکی، به‌ویژه در ارزیابی سیلاب‌ها و برآورد دبی طراحی برای دوره‌های بازگشت مختلف، داده‌های پرت نقش مهمی در دقت برآورد دارند. نتایج این پژوهش نشان

نویسنده اول: جمع‌آوری و تجزیه و تحلیل داده‌ها، نگارش متن مقاله

نویسنده دوم و سوم: راهنمایی، ویرایش و بازبینی مقاله، کنترل نتایج

نویسنده چهارم: مشاوره و بازبینی متن مقاله

اصول اخلاقی

نویسندگان اصول اخلاقی را در انجام و انتشار این اثر عملی رعایت نموده‌اند و این موضوع مورد تأیید همه آن‌ها می‌باشد.

منابع

- Ahmadi, M. and Sarmad, M., 2010. Detecting outliers in normal data using modified Z-scores. *Journal of Statistical Sciences*, 3(2), pp. 119–139.
- Bačová Mitková, V., Pekárová, P., Halmová, D. and Miklášek, P., 2021. The use of a uniform technique for harmonization and generalization in assessing the flood discharge frequencies of long return period floods in the Danube River Basin. *Water*, 13(10), p. 1337. <https://doi.org/10.3390/w13101337>.
- Bhat, M.S., Alam, A., Ahmad, B., Kotlia, B.S., Farooq, H., Taloor, A.K. and Ahmad, S., 2019. Flood frequency analysis of River Jhelum in Kashmir Basin. *Quaternary International*, 507, pp. 288–294. <https://doi.org/10.1016/j.quaint.2018.09.039>
- Dadkhah, K. and Samadi Tudar, E., 2018. Robust analysis of variance based on permutation distribution of trimmed mean. *Journal of Statistical Sciences*, 12(1), pp. 119–141. [In Persian].
- Dehghan, S. and Faridrohani, M., 2022. Multivariate outlier detection based on depth-based outlyingness function. *Journal of Statistical Sciences*, 15(2), pp. 443–462. [In Persian].
- Doocy, S., Daniels, A., Murray, S. and Kirsch, T.D., 2013. The human impact of floods: A historical review of events 1980–2009 and systematic literature review. *PLoS Currents*, 5. <https://doi.org/10.1371/currents.dis.f4deb457904936b07c09daa98ce8171a>
- Griffin, A., Shaw, L. and Stewart, E., 2018. Technical note: Approximate Bayesian computation to improve long-return flood estimates using historical data. *Hydrology and Earth System Sciences Discussions*. <https://doi.org/10.5194/hess-23-3057-2019>.
- Hamzah, F.M., Tajudin, H. and Jaafar, O., 2021. A comparative flood frequency analysis of highflow between annual maximum and partial duration series at Sungai Langat Basin. *Sains Malaysiana*, 50, pp. 1843–1856. <https://doi.org/10.17576/jsm-2021-5007-02>.
- Heydarpour, B., Sagafian, B., Shamsaei, A. and Goliyan, S., 2015. Evaluation of the impact of incorporating extraordinary floods in flood

داده‌ها دارد، تأثیر قابل‌توجهی بر تحلیل‌های آماری و مدل‌سازی‌های هیدرولوژیکی می‌گذارد. در صورت حذف داده دبی سال ۱۳۹۴ به عنوان داده پرت، برای دوره بازگشت ۵۰۰ ساله دبی طراحی به‌طور قابل‌توجهی و به میزان ۶۴/۵۸ درصد کاهش یافت. این کاهش در دبی طراحی نشان‌دهنده این است که داده‌های پرت می‌توانند بر برآورد دبی‌های طراحی برای دوره بازگشت‌های طولانی مدت تأثیر زیادی بگذارند و ممکن است باعث کم برآورد شدن ریسک سیلاب‌های شدید در این دوره شوند. با این حال، اگر داده دبی مربوط به سال ۱۳۹۴ به‌عنوان یک داده کم احتمال در محاسبات گنجانده شود، دبی طراحی باید به‌طور معقول‌تری برای دوره بازگشت ۱۰۰ ساله محاسبه گردد. داده پرت به‌عنوان بخشی از توزیع طبیعی دبی‌ها در منطقه است و می‌تواند به بهبود دقت پیش‌بینی دبی‌های طراحی و همچنین طراحی سازه‌های هیدرولوژیکی کمک کند. گنجاندن داده‌های پرت در مدل‌سازی‌ها از این جهت مهم است که می‌تواند تأثیرات سیلاب‌های بزرگ و کم احتمال را بهتر مدل‌سازی کند و در نتیجه، طراحی‌های مقاوم‌تری برای زیرساخت‌ها فراهم آورد.

این یافته‌ها تأکید می‌کنند که مدیریت داده‌های پرت و تصمیم‌گیری صحیح در مورد گنجاندن یا حذف آن‌ها از مجموعه داده‌ها، به‌ویژه در تحلیل‌های مربوط به سیلاب‌ها و دبی‌های طراحی، می‌تواند تأثیرات زیادی بر دقت پیش‌بینی‌ها و طراحی‌های هیدرولوژیکی داشته باشد. انتخاب روش مناسب برای مدیریت این داده‌ها باید با توجه به ماهیت داده‌ها و هدف تحلیل انجام شود تا از خطاهای قابل توجه در پیش‌بینی‌ها جلوگیری شود.

تقدیر و تشکر

نویسندگان از سازمان هواشناسی و شرکت آب منطقه‌ای استان ایلام به خاطر در اختیار قرار دادن داده و اطلاعات برای انجام این پژوهش تشکر می‌نمایند.

دسترسی به داده‌ها

بخشی از داده‌های این تحقیق از سازمان هواشناسی بخشی دیگر از وزارت نیرو دریافت شده است.

مشارکت نویسندگان

نحوه و میزان مشارکت نویسندگان در انجام این پژوهش به‌صورت زیر است:

- and Using Geographic Information System (GIS). *Iran-Watershed Management Science & Engineering*. Vol. 8, No. 25. [In Persian].
- 23- Ramasamy, M. and Nagan, S., 2022. Analysis of flood frequency using plotting position methods and Gumbel's method: Vaigai River basin, Southern India. *Global NEST Journal*, 24(4), pp. 671–680.
 - 24- Rasekh, A., Afshar, A. and Afshar, M.H., 2010. Risk-cost optimization of hydraulic structures: Methodology and case study. *Water Resources Management*, 24(11), pp. 2833–2851. <https://doi.org/10.1007/s11269-010-9582-3>.
 - 25- Pipiras, V., 2020. Pitfalls of data-driven peaks-over-threshold analysis: perspectives from extreme ship motions. *Probabilistic Engineering Mechanics*, 60, Art. 103053. [doi: 10.1016/j.probengmech.2020.103053](https://doi.org/10.1016/j.probengmech.2020.103053).
 - 26- Peruri VenkataAnusha, C., Anuradha, C., Chandra Murty, P.S.R. and Chebrolu, C.S.K., 2019. Detecting outliers in high dimensional data sets using Z-Score methodology. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 9(1), pp. 5067–5070. <https://doi.org/10.35940/ijitee.A3910.119119>.
 - 27- Rostampour, M., 2022. Comparison of outlier detection methods and their impact on rangeland measurement and assessment studies. *Iranian Journal of Natural Resources*, 75(4). [In Persian].
 - 28- Rousseeuw, P.J. and Leroy, A.M., 1987. *Robust regression and outlier detection*. Wiley. ISBN-10: 0471488550, ISBN-13: 978-0471488554
 - 29- Saghafian, B., Ghasemi, A. and Galian, S., 2013. Frequency analysis of flood discharge based on rainfall-runoff model simulation and statistical distributions. *Iranian Journal of Water and Soil Research*, 44(1), pp. 21–32. [In Persian].
 - 30- Strupczewski, W.G., Kochanek, K. and Bogdanowicz, E., 2014. On return period of the largest historical flood. *Journal of Geoscience and Environment Protection*, 2, pp. 144–152.
 - 31- Turhan, E. and Değerli, S., 2022. A comparative study of probability distribution models for flood discharge estimation: Case of Kravga Bridge, Turkey. *Geofizika*, 39, pp. 57–70.
 - 32- Wu, Y., Zhu, H., Luo, T., Liu, J. and Zhang, Y., 2023. The analysis on the influence of heavy rain and flood on the design flood of Liqingdian Hydrological Station. *Desalination and Water Treatment*, 314, pp. 368–375.
 - frequency analysis. *Watershed Engineering and Management*, 7(3), pp. 317–330. [In Persian].
 - 10- Houshmand, J., Delghandi, M. and Seir Kaboli, J., 2005. Karoon's river's water qualitative paneling by NSFQI index & GIS. In: *Proceedings of the 2nd Congress on Environmental Engineering*, Tehran University. [In Persian].
 - 11- Khwairakpam, R.S., 2024. Comparative analysis of flood estimation using Log-Pearson Type III and Gumbel Max models in the Cauvery River, India. *International Journal of Innovative Science and Research Technology*, 9(4).
 - 12- Kiani, R. and Montazeri, M., 2015. A review of outlier detection methods. *International Conference on Research in Science and Technology*. [In Persian].
 - 13- Kite, G.W., 1977. *Frequency and risk analysis in hydrology*. Water Resources Publications. ISBN-10: 0918334427, ISBN-13: 978-0918334420.
 - 14- Lefebvre, M., 2022. Estimation of the return periods in hydrology. *Advances in Environmental and Engineering Research*, 3(4), pp. 45–58. <https://ojs.bilpublishing.com/index.php/aeer>.
 - 15- Machado, M.J., Botero, B.A., López, J., Francés, F., Díez-Herrero, A. and Benito, G., 2015. Flood frequency analysis of historical flood data under stationary and non-stationary modelling. *Hydrology and Earth System Sciences*, 19(6), pp. 2561–2576. [doi: 10.5194/hess-19-2561-2015](https://doi.org/10.5194/hess-19-2561-2015).
 - 16- Mahdavi, M., 2005. *Applied Hydrology*. Vol.2. Tehran Univercitr Prees. 427p.
 - 17- Mahmoodi, Z. and Mesbahi, M., 2016. Estimation of flood discharge with different return periods using superior statistical distributions (Case study: Kor and Sivand Watershed). *National Conference on Water and Hydraulic Structures*, Dezful, Iran. . [In Persian].
 - 18- Mangukiya, N.K. and Sharma, A., 2024. Alternate pathway for regional flood frequency analysis in data-sparse regions. *Journal of Hydrology*, 629. <https://doi.org/10.1016/j.jhydrol.2023.130635>
 - 19- Nathan, R.J. and McMahon, T.A., 1990. Identification of homogeneous regions for the purpose of regionalization. *Journal of Hydrology*, 121(1-4), pp. 217–238. [https://doi.org/10.1016/0022-1694\(90\)90233-N](https://doi.org/10.1016/0022-1694(90)90233-N).
 - 20- Okoli, K., Breinl, K., Brandimarte, L., Botto, A., Volpi, E. and Di Baldassarre, G., 2018. Model averaging versus model selection: Estimating design floods with uncertain river flow data. *Hydrological Sciences Journal*, 63(13–14), pp. 1913–1926. <https://doi.org/10.1080/02626667.2018.155733>.
 - 21- Okoli, K., Mazzoleni, M., Breinl, K. and Di Baldassarre, G., 2019. A systematic comparison of statistical and hydrological methods for design flood estimation. *Hydrology Research*. <https://doi.org/10.2166/nh.2019.030>
 - 22- Ramazani, A., Ahmadi Moghadam, M., and Jaafari, M.R., 2014. Water Quality Zoning of Godar-Khosh River Based on the NSFQI Index