Iranian Journal of Rainwater Catchment Systems
Online ISSN: 1531-2783 Print ISSN: 5970-2423

Calibration of ERA5 daily precipitation using MLP, D-Tree, and KNN algorithms in Razavi Khorasan province

Majid Rajabi Jaghargh[1], Seyed Mohammad Mousavi Baygi[*], Seyed Alireza Araghi[3], Hadi Jabari Noghabi[4]

1. Ph.D. Student, Department of Water Science and Engineering, Faculty of Agriculture, Ferdowsi University of Mashhad, Mashhad, Iran. Email: magidrajabijaghargh@mail.um.ac.ir
2. Professor, Department of Water Science and Engineering, Faculty of Agriculture, Ferdowsi University of Mashhad, Mashhad, Iran. Email: mousavib@um.ac.ir
3. Assistant Professor, Department of Water Science and Engineering, Faculty of Agriculture, Ferdowsi University of Mashhad, Mashhad, Iran. Email: a.araghi@um.ac.ir
4. Associate Professor, Department of Statistics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Mashhad, Iran. Email: jabbarimh@um.ac.ir

ARTICLE INFO

Article type:
Research Paper

Article history
Received: 27 November 2023
Revised: 11 January 2024
Accepted: 14 January 2024
Published online: 06 January 2024

Keywords:
Calibration, database, statistical indicators, machine learning

ABSTRACT

Satellite products are the only available data source with adequate spatial coverage, however, their data do not match the observed values and have biases, although this discrepancy cannot be fixed precisely, however, a solution to reduce the bias is data recalibration. Currently, machine learning techniques are used to improve the accuracy of forecasting various types of weather phenomena, so regression solving such problems through methods based on machine learning and deep learning is very efficient. The daily precipitation of 19 rain gauge stations of the Ministry of Energy between 2010 and 2021 was extracted and compared to the average values of their corresponding daily precipitation pixels in the ERA5 database. To measure the data, three algorithms D-Tree, KNN, and MLP were used. The range of changes of correlation coefficient in MLP, D-Tree, and KNN is equal to [0.87, 0.98], [0.75, 0.97], and [0.4, 0.87], respectively. In addition, the range of changes for RMSE in MLP varies from 0.7 to 2.4 mm per day, and these changes for D-Tree and KNN are calculated between 0.8 to 2.2 and 1.2 to 2.5, respectively. In 75% of stations, RMSE in MLP, D-Tree, and KNN algorithms is less than 1.5, 1.9, and 2.2 mm per day, respectively. The range of bias changes in MLP is [0.18, -0.6 mm per day] and this range of changes for D-Tree and KNN is respectively [0.16, 0.5 mm per day] and [0.6, -0.8 mm per day] have been calculated. The bias of corrected data and observed values in MLP, D-Tree, and KNN algorithms for the middle of the stations is -0.09, -0.11, and -0.16 mm per day, respectively. The evaluation of the performance of three machine learning algorithms (MLP, D-Tree, and KNN) in correcting the daily precipitation of the ERA5 database and the comparison of CC, RMSE, and bias statistical indices for the reproduced data compared to ground values showed that in all three statistical indices, the MLP algorithm works better than the others and has good accuracy for correcting the daily precipitation.

Citation: Rajabi Jaghargh, M., Mousavi Baygi, S. M, Araghi, S.A, & Jabari Noghabi, H. (2024). Calibration of ERA5 daily precipitation using MLP, D-Tree, and KNN algorithms in Razavi Khorasan province. *Iranian Journal of Rainwater Catchment Systems*, 12(1), 129-147.
DOR: 20.1001.1.24235970.1403.12.1.8.1

Publisher: Iranian Rainwater Catchment Systems Association

***Corresponding author:** Seyed Mohammad Mousavi Baygi
Address: Department of Water Science and Engineering, Ferdowsi University of Mashhad, Mashhad, Iran.
Tel: +989153167311
Email: mousavib@um.ac.ir

Iranian Journal of Rainwater Catchment Systems
Online ISSN: 1531-2783 Print ISSN: 5970-2423

Calibration of ERA5 daily precipitation using MLP, D-Tree, and KNN algorithms in Razavi Khorasan province

Majid Rajabi Jaghargh[1], Seyed Mohammad Mousavi Baygi[*2], Seyed Alireza Araghi[3], Hadi Jabari Noghabi[4]

1. Ph.D. Student, Department of Water Science and Engineering, Faculty of Agriculture, Ferdowsi University of Mashhad, Mashhad, Iran. Email: magidrajabijaghargh@mail.um.ac.ir
2. Professor, Department of Water Science and Engineering, Faculty of Agriculture, Ferdowsi University of Mashhad, Mashhad, Iran. Email: mousavib@um.ac.ir
3. Assistant Professor, Department of Water Science and Engineering, Faculty of Agriculture, Ferdowsi University of Mashhad, Mashhad, Iran. Email: a.araghi@um.ac.ir
4. Associate Professor, Department of Statistics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Mashhad, Iran. Email: jabbarimh@um.ac.ir

EXTENDED ABSTRACT

Introduction: Spatial and temporal changes in precipitation significantly affect access to water for various human needs and environmental conservation. Although ground-based rain gauge systems are the most reliable tools for monitoring changes in precipitation at a point scale, various factors such as topography, remote locations, and budget constraints limit their geographical use. Therefore, ground-based rain gauge networks in many parts of the world are often spatially sparse and have weaknesses in spatial coverage. Satellite products are the only available data source with global coverage; however, there is a mismatch between them and ground observations. The values provided by sensors and data platforms cannot accurately estimate precipitation data due to multiple issues. These problems are not easily solvable, however, one solution to reduce ambiguities is to calibrate the estimated data. Currently, machine learning techniques are being employed to improve the accuracy of predicting various weather phenomena. Therefore, solving regression problems of this nature through machine learning and deep learning methods is not only possible but also very efficient. To this end, this study aimed to calibrate daily precipitation values in the ERA5 database by fitting these values against daily precipitation from 19 rain gauge stations of the Ministry of Energy in the Khorasan Razavi province, using three algorithms: D-Tree, K-NN, and MLP.

Methodology: Khorasan-Razavi province is located in the northeast of Iran with an area of about 117,000 km^2. This province contains arid and semi-arid areas with a wide range of temperature changes and precipitation patterns. Half of its area is made up of mountainous areas and the other half is plain and low-altitude areas. The climate of this province is cold dry and its average annual precipitation is about 254 mm. The southern regions of this province have less than 150 mm of precipitation per year, and the central and northwestern regions receive more than 320 mm. In this study, we used three popular and widely used algorithms, decision trees, K-nearest neighbors, and artificial neural networks for calibrating precipitation values. Daily precipitation values from 19 rain gauge stations of the Ministry of Energy have been extracted since 2010 and compared to the average daily precipitation pixel values from the ERA5 database. Using the above algorithms, terrestrial precipitation values were estimated based on ERA5 data, and correlation coefficient, RMSE, and bias indices were extracted and compared for each of the three algorithms.

Results and Discussion: Performance evaluation of machine learning algorithms (MLP, D-Tree, KNN) in correcting daily precipitation from ERA5 database and comparing them with statistical indices CC, RMSE, and bias for the reproduced data relative to ground values showed that in all three statistical indices, the MLP algorithm performed better than the other two algorithms and had suitable accuracy for correcting daily precipitation from ERA5. The range of changes in correlation coefficient in MLP, D-Tree, and KNN is [0.87, 0.98], [0.75, 0.97], and [0.4, 0.87], respectively, and these changes for RMSE in MLP are between 0.7 and 2.4 mm per days. These

***Corresponding author:** Seyed Mohammad Mousavi Baygi
Address: Department of Water Science and Engineering, Ferdowsi University of Mashhad, Mashhad, Iran.
Tel: +989153167311
Email: mousavib@um.ac.ir

[DOR: 20.1001.1.24235970.1403.12.1.8.1]

ISSN (Online): 2783-1531 - ISSN (Print): 2423-5970

changes for D-Tree and KNN are calculated between 0.8 to 2.2 and 1.2 to 2.5, respectively, in 75 stations, the RMSE in MLP, D-Tree, and KNN algorithms is less than 5.5, 1.9, and 2.2 mm per day. The range of bias changes in MLP [0.18, -0.6 mm d^{-1}] and this range of changes for D-Tree and KNN is respectively calculated as [0.16, 0.5 mm d^{-1}] and [0.6, -0.8 mm d^{-1}]. The corrected and observed bias in MLP, D-Tree, and KNN algorithms for the middle of the stations are equal to -0.09, -0.11, and -0.16 mm per day, respectively.

Conclusion: With the advent of advanced technologies and the presence of satellites and associated databases in the field of meteorological science, the time series of satellite precipitation values have become long enough to analyze their applicability for water resources management. Although these data are cheap and easily accessible, they do not have enough accuracy and they need to be calibrated with ground values. This study was shown that MLP performed better in calibrating daily precipitation values in ERA5 compared to the other two algorithms, as it effectively increased the correlation coefficient and desirably reduced RMSE and bias. For further informations, it is suggested to use deep learning techniques for calibrating monthly and annual precipitation values, as well as to investigate and analyze the accuracy of microscaling and correction of satellite precipitation data in the Google Earth Engine system.

Ethical Considerations

Data availability statement: The datasets are available upon a reasonable request to the corresponding author.
Funding: This research has been extracted from a doctoral thesis under the support of Ferdowsi University of Mashhad.
Authors' contribution: Majid Rajabi Jaghargh prepared the first draft and it was edited by Seyed Mohammad Mousavi Baygi, Seyed Alireza Araghi, and Hadi Jabari Noghabi.
Conflicts of interest: The authors of this article declared no conflict of interest regarding the authorship or publication of this article.
Acknowledgment: Khorasan Razavi Regional Water Company is thanked for providing the required statistics and information.

[DOR: 20.1001.1.24235970.1403.12.1.8.1]

Iranian Journal of Rainwater Catchment Systems
مجله علمی سامانه‌های سطوح آبگیر باران ایران
شاپا الکترونیکی: ۲۷۸۳-۱۵۳۱ شاپا چاپی: ۲۴۲۳-۵۹۷۰

واسنجی بارش روزانه ERA5 با استفاده از الگوریتم‌های MLP، D-Tree و KNN در استان خراسان‌رضوی

مجید رجبی جاغرق[۱]، سید محمد موسوی بایگی[۲*]، سید علیرضا عراقی[۳]، هادی جباری نوقابی[۴]

۱. دانشجوی دکتری، گروه علوم و مهندسی آب، دانشکده کشاورزی، دانشگاه فردوسی مشهد، مشهد، ایران. magidrajabijaghargh@mail.um.ac.ir
۲. استاد، گروه علوم و مهندسی آب، دانشکده کشاورزی، دانشگاه فردوسی مشهد، مشهد، ایران. mousavib@um.ac.ir
۳. استادیار، گروه علوم ومهندسی آب، دانشکده کشاورزی، دانشگاه فردوسی مشهد، مشهد، ایران. a.araghi@um.ac.ir
۴. دانشیار، گروه آمار، دانشکده علوم ریاضی، دانشگاه فردوسی مشهد، مشهد، ایران. jabbarimh@um.ac.ir

مشخصات مقاله

نوع مقاله:
پژوهشی

تاریخچه مقاله
دریافت: ۶ آذر ۱۴۰۲
بازنگری: ۲۱ دی ۱۴۰۲
پذیرش: ۲۴ دی ۱۴۰۲
انتشار برخط: ۱۶ خرداد ۱۴۰۳

واژه‌های کلیدی:
پایگاه داده، شاخص‌های آماری، یادگیری ماشین، واسنجی

چکیده

محصولات ماهواره‌ای تنها منبع داده موجود با پوشش فضایی مناسب است، با این وجود، داده‌های آن‌ها بر مقادیر مشاهداتی منطبق نبوده و دارای انحراف است، هرچند این عدم تطابق به‌طور دقیق قابل رفع نیست، با این حال، یک راه‌حل کاهش سوگیری، واسنجی داده‌هاست. در حال حاضر، تکنیک‌های یادگیری ماشین برای بهبود دقت پیش‌بینی انواع مختلف پدیده‌های آب و هوایی به کار گرفته می‌شوند، لذا حل رگرسیونی مسائلی از این قبیل از طریق روش‌های مبتنی بر یادگیری ماشین و آموزش عمیق بسیار کارآمد است. باراش روزانه ۱۹ ایستگاه باران‌سنج ثبات وزارت نیرو بین سال‌های ۲۰۱۰ تا ۲۰۲۱ میلادی استخراج شد و در مقابل مقادیر متوسط پیکسل‌های بارش روزانه متناظر آن‌ها در پایگاه داده ERA5 قرارگرفت. به‌منظور واسنجی داده‌ها، از سه الگوریتم D-Tree، KNN و MLP استفاده شد. دامنه تغییرات ضریب همبستگی در MLP، D-Tree و KNN به‌ترتیب برابر [0.87, 0.98]، [0.75, 0.97] و [0.4, 0.87] است. هم‌چنین این دامنه تغییرات برای RMSE در MLP بین ۰/۷ تا ۲/۴ میلی‌متر در روز متغیر بوده و این تغییرات برای D-Tree و KNN به‌ترتیب بین ۰/۸ تا ۲/۲ و ۱/۲ تا ۲/۵ محاسبه شده‌اند. در ۷۵ درصد ایستگاه‌ها RMSE در الگوریتم‌های MLP، D-Tree وKNN به‌ترتیب کم‌تر از ۱/۵، ۱/۹ و ۲/۲ میلی‌متر در روز است. دامنه تغییرات سوگیری در MLP، [۰/۱۸، ۰/۶- میلی‌متر در روز] بوده و این دامنه تغییرات برای D-Tree و KNN به‌ترتیب [۰/۱۶، ۰/۵ میلی‌متر در روز] و [۰/۶، ۰/۸- میلی‌متر در روز] محاسبه شده‌اند. سوگیری داده‌های اصلاحی و مقادیر مشاهده شده، در الگوریتم‌های MLP، D-Tree وKNN برای میانه ایستگاه‌ها به‌ترتیب برابر ۰/۰۹-، ۰/۱۱- و ۰/۱۶- میلی‌متر در روز است. ارزیابی عملکرد سه الگوریتم یادگیری ماشین (MLP، D-Tree و KNN) در تصحیح بارش روزانه پایگاه داده ERA5 و مقایسه شاخص‌های آماری CC، RMSE و سوگیری برای داده‌های بازتولید شده نسبت به مقادیر زمینی نشان داد که در هر سه شاخص آماری الگوریتم MLP نسبت به دوالگوی دیگر بهتر عمل نموده و از دقت مناسبی برای تصحیح بارش روزانه برخوردار است.

استناد: رجبی جاغرق، مجید، موسوی بایگی، سید محمد، عراقی، سید علیرضا و جباری نوقابی، هادی (۱۴۰۳). واسنجی بارش روزانه ERA5 با استفاده از الگوریتم‌های MLP، D-Tree و KNN در استان خراسان‌رضوی. سامانه‌های *سطوح آبگیر باران*، ۱۲(۱)، ۱۲۹-۱۴۷.

DOR: 20.1001.1.24235970.1403.12.1.8.1

ناشر: انجمن علمی سیستم‌های سطوح آبگیر باران ایران

* **نویسنده مسئول:** سید محمد موسوی بایگی
نشانی: گروه علوم ومهندسی آب، دانشکده کشاورزی، دانشگاه فردوسی مشهد، مشهد، ایران
تلفن: ۰۹۱۵۳۱۶۷۳۱۱
پست الکترونیکی: mousavib@um.ac.ir

مقدمه

بارش یکی از متغیرهای کلیدی در چرخه آب و اصلی‌ترین داده در مدل‌سازی‌های هیدرولوژیکی محسوب می‌شود (Yuan et al., 2018, Beck et al., 2017، Liu et al., 2017). بنابراین داده‌های دقیق و قابل اعتماد بارش نه تنها برای بهبود دقت شبیه‌سازی‌های هیدرولوژیکی بسیار مهم هستند (Sun et al., 2018، Prakash et al., 2015, Pena et al., 2013)، بلکه برای فهم بهتر از منابع آب و فرآیندهای آب و هواشناسی ضروری هستند (Ma et al., 2015). امروزه مقادیر بارش‌های تخمینی مبتنی بر اطلاعات ماهواره‌ای، به‌عنوان جایگزینی مناسب در مطالعات مرتبط با علوم آب و محیط زیست، به‌طور فزاینده‌ای در دسترس عموم قرار گرفته است (Li et al., 2018)، اگرچه این محصولات دارای یک پوشش فضایی وسیع و وضوح مکانی و زمانی مناسب هستند (Zubieta et al., 2015)، اما از آن‌جا که الگوریتم‌های تخمین بارش ماهواره‌ای، به‌صورت جهانی تنظیم شده‌اند و امکان نادیده گرفتن عوامل تأثیرگذار محلی، در تخمین بارش بسیار محتمل است، لذا عدم تطابق بین داده‌های ماهواره‌ای و زمینی دور از انتظار نیست، بنابراین ضروری است تا نسبت به اصلاح و واسنجی این مقادیر با استفاده از روش‌های موجود اقدام شود. پژوهشگران برای اصلاح داده‌های تخمینی، بسته به میزان دقت مورد نیاز خود از روش‌های مختلفی استفاده می‌کنند. برای نمونه، Chen et al. (2013) روش‌های تصحیح سوگیری را به دو دسته طبقه‌بندی نمودند: ۱) رویکردهای متکی بر میانگین، که معمولاً از عوامل تصحیح ماهانه بر اساس نسبت بین بارش مبتنی بر شبکه و مقادیر مشاهده شده استفاده می‌کنند، ۲) رویکردهای مبتنی بر توزیع، که معمولاً در آن‌ها از توابع احتمالی استفاده می‌شود، مانند تابع توزیع تجمعی (CDF) (Yang et al., 2010, Piani et al., 2010، Amengual et al., 2012، Vrac et al., 2016، Hamill et al., 2018، Guo et al., 2018). رویکردهای متکی بر میانگین، روشی به نسبت ساده و آسان برای پیاده‌سازی در مقایسه با روش تابع توزیع تجمعی است، لذا در مطالعاتی از این قبیل با اقبال همراه بوده است (Themeßl et al., 2011، Chen et al., 2013).

وجود خطاهای تصادفی، برای تخمین بارش ماهواره‌ای در مقیاس روزانه، باعث سوگیری این داده‌ها نسبت به مقادیر بارش ایستگاه‌های زمینی می‌شود، لذا برای تصحیح این ناسازگاری فضایی، محققین اقدام به ایجاد رابطه بین بارش و متغیرهای مکانی نموده‌اند. ژانگ (Zhang et al. , 2019a)، کل منطقه حوضه آبریز رودخانه لنکانگ–مکونگ را با توجه به نتایج تجزیه و تحلیل REOF[1] برای بارش زمینی و ماهواره‌ای به هفت زیرمنطقه تقسیم کرد و دریافت، کاهش مقیاس، باعث ارائه نتایج بهتر می‌شود. Zhang et al. (2019b) عملکرد و کاربرد هیدرولوژیکی TMPA[2] را ارزیابی نمودند. آن‌ها منطقه مورد مطالعه را به چهار زیرحوضه تقسیم کردند و مدل تنظیم جغرافیایی را برای تصحیح داده‌های TMPA بر اساس چند ضلعی‌های تیسن[3] در هر زیرحوضه به کار بردند و نتایج رضایت‌بخشی را به‌دست آوردند. Heredia et al. (2018) برای انطباق فضایی بارش، رویکرد تابع توزیع تجمعی[4] (CDF) را برای تصحیح مقیاس نقطه‌ای و اصلاح شبکه، با تقسیم ناحیه مورد نظر، با استفاده از چندضلعی‌های تیسن پیشنهاد کرد. Guo et al. (2018) معتقد است، بارش تخمینی توسط ماهواره، به شدت بارش وابسته بوده وبرای اثبات ادعای خود، شش حوضه آبریز درکشور چین را که از نظرویژگی‌هایی مانند اقلیم، عرض جغرافیایی، طول جغرافیایی و ارتفاع متفاوت بودند را انتخاب نمود و نسبت به اعتبارسنجی بارش در آن‌ها اقدام کرد، او دریافت که روش CDF را باید در یک منطقه خاص به‌کار بست تا بتواند به‌طور موثری ویژگی‌های بارش را در رژیم‌های مختلف اقلیمی منعکس نماید.

دانستن مقدار بارش در یک شبکه متراکم فضایی و برای یک دوره زمانی طولانی می‌تواند در حل انواع مسائل مهندسی، مثمر ثمر باشد (Blöschl et al., 2019)، اگر چه، داده‌های شبکه بارانسنج زمینی ازدقت بالایی برخوردارند، اما نگهداری از چنین شبکه‌ای با تراکم فضایی بالا و برای مدت زمان طولانی بسیار هزینه‌بر است (Mega et al., 2019, Tang et al., 2022). لذا باید به دنبال روش‌های جایگزین گشت که قادر باشد هم از نظر زمانی و هم به لحاظ مکانی مقادیر بارش را با دقت مطلوبی ارائه نماید.

با ظهور فناوری‌های پیشرفته و حضور ماهواره‌ها و پایگاه‌های داده مرتبط درحوزه علم هواشناسی، سری‌های زمانی مقادیر بارش ماهواره‌ای به اندازه کافی طولانی شده‌اند تا قابلیت استفاده از آن‌ها را برای مدیریت منابع آب، تجزیه و تحلیل نمود، هرچند این داده‌های تولیدی ارزان و سهل‌الوصول هستند اما از دقت کافی برخوردار نیستند و برای استفاده از آن‌ها نیاز است با مقادیر زمینی واسنجی شوند (Mega et al., 2019, Tang et al., 2022).

[1] Rotated Empirical Orthogonal Function
[2] TRMM Multi-Satellite Precipitation Analysis
[3] Thiessen
[4] Cumulative Distribution Function

[DOR: 20.1001.1.24235970.1403.12.1.8.1]

با ادغام محصولات بارش ماهواره‌ای شبکه‌بندی شده و اندازه‌گیری‌های زمینی، این امکان وجود دارد به داده‌هایی با دقت و کیفیت بالاتر دست یافت، که نه تنها دقیق‌تر از داده‌های ماهواره‌ای است بلکه به‌طور هم‌زمان، فضا را با تراکم بسیار بالاتر در مقایسه با اندازه‌گیری‌های زمینی پوشش می‌دهد. این ادغام، عملاً حل رگرسیونی مسئله در یک محیط مکانی است، این روش‌ها معمولاً تحت عنوان کاهش مقیاس نیز معرفی می‌شوند و نوع خاصی از درون‌یابی فضایی هستند. مسئله اخیر در زمینه‌های مختلفی قابل مشاهده است و می‌توان انواع روش‌های مرتبط با کاهش مقیاس بارش را در Hu et al. (2019) و Abdollahipour et al. (2022) به‌طور جامع‌تری یافت.

درون‌یابی فضایی بارش با ادغام محصولات بارش ماهواره‌ای و اندازه‌گیری‌های زمینی در مقیاس‌های زمانی و مکانی مختلف و با استفاده از انواع الگوریتم‌های رگرسیونی از جمله الگوریتم‌های یادگیری ماشین توسط محققان مختلفی انجام شده است. خلاصه اطلاعات روش‌شناختی برخی از این مطالعات در جدول (۱) خلاصه شده است.

جدول ۱- خلاصه اطلاعات روش‌شناختی الگویتم‌های یادگیری ماشین

Table 1- Summary of methodological information of machine learning algorithms

Study	Time Scale	Spatial Scale	Algorithms
He et al. 2016	Hourly	South-western, central, north-eastern and south-eastern United States	Random forests
Meyer et al. 2016	Daily	Germany	Random forests, artificial neural networks, support vector regression
Tao et al. 2016	Daily	Central United States	Deep learning
Yang et al. 2016	Daily	Chile	Quantile mapping
Baez-Villanueva et al. 2020	Daily	Chile	Random forests
Chen et al. 2020	Daily	Dallas–Fort Worth in the United States	Deep learning
Chen et al. 2020	Daily	Xijiang basin in China	Geographically weighted ridge regression
Rata et al. 2020	Annual	Chéliff watershed in Algeria	Kriging
Chen et al. 2021	Monthly	Sichuan Province in China	Artificial neural networks, geographically weighted regression, kriging, random forests
Nguyen et al. 2021	Daily	South Korea	Random forests
Shen and Yong 2021	Annual	China	Gradient boosting decision trees, random forests, support vector regression
Zhang et al. 2021	Daily	China	Artificial neural networks, extreme learning machines, random forests, support vector regression
Chen et al. 2021	Daily	Coastal mountain region in the western United States	Deep learning
Fernandez Palomino et al. 2022	Daily	Ecuador and Peru	Random forests
Lin et al. 2022	Daily	Three Gorges Reservoir area in China	Adaptive boosting decision trees, decision trees, random forests
Yang et al. 2022	Daily	Kelantan river basin in Malaysia	Deep learning
Zandi et al. 2022	Monthly	Alborz and Zagros mountain ranges in Iran	Artificial neural networks, locally weighted linear regression, random forests, stacked generalization, support vector regression
Militino et al. 2023	Daily	Navarre in Spain	K-nearest neighbors, random forests, artificial neural networks

یادگیری ماشین روشی است که در سال‌های اخیر به طور گسترده مورد استفاده قرار گرفته است، با پیشرفت فن‌آوری، این روش به یک روش عمومی برای طبقه‌بندی، و هم‌چنین برای مشکلات تشخیص تغییرات و ناهنجاری در علوم زمین تبدیل شده است (Gómez-Chova et al., 2015, Muhlbauer et al., 2014). برای هواشناسی، که شاخه‌ای از علوم زمین است، روش‌های یادگیری ماشین نیز با موفقیت به‌کار گرفته شده، به‌عنوان مثال، شبکه‌های عصبی مصنوعی (ANN) برای بهبود تخمین بارش (Pan et al., 2019)، شناسایی مرحله ENSO و پیش‌بینی تأثیر آن از جمله این موارد هستند (Toms et al., 2020).

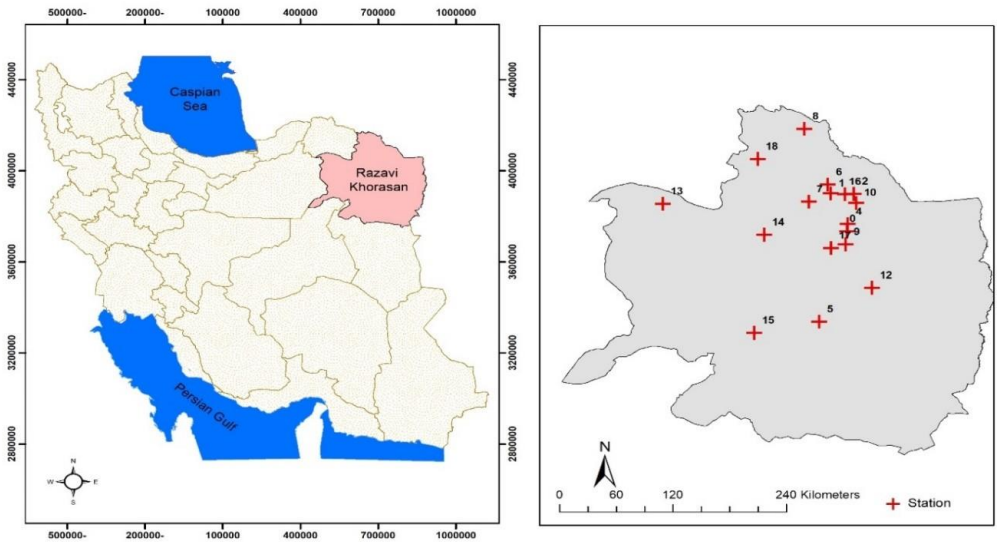
[DOR: 20.1001.1.24235970.1403.12.1.8.1]

انتخاب استان خراسان‌رضوی به‌عنوان منطقه هدف دراین تحلیل، به واسطه دسترسی به گراف‌های باران‌سنج‌های ثبات وزارت نیرو برای تحلیل‌های آماری بود. با توجه به این‌که مقادیر بارش سنجنده‌ها و پایگاه‌های داده می‌تواند در آینده‌ای نزدیک جایگزین شبکه ایستگاه‌های زمینی شود، بنابراین استفاده از محصولات آن‌ها به سرعت درحال افزایش است، اما مقادیر بارش تخمینی توسط ماهواره‌ها و پایگاه‌های داده بر مقادیر مشاهداتی منطبق نبوده لذا واسنجی داده‌های تخمینی را ضروری می‌نماید، لذا باید تا نسبت به ارزیابی دقت داده‌های ماهواره‌ای و واسنجی آن‌ها با استفاده ازروش‌های نوین ازجمله یادگیری ماشین اقدام شود تا تطابق داده‌ها حداکثر و میزان خطاها به حداقل ممکن برسد. آن‌چه نقطه قوت و وجه تمایز این پژوهش است شامل استفاده از سه الگوریتم یادگیری ماشین (درخت تصمیم، K نزدیک‌ترین همسایگان، پرسپترون چندلایه) و انتخاب بهترین برآوردگر، برای واسنجی و کاهش سوگیری مقادیر بارش روزانه پایگاه ERA5 در استان خراسان‌رضوی است.

مواد و روش تحقیق

محدوده مورد مطالعه

استان خراسان‌رضوی در شمال شرقی ایران واقع شده و دارای مساحتی درحدود ۱۱۷ هزارکیلومترمربع است، این استان، مناطق خشک و نیمه خشک با طیف گسترده‌ای از تغییرات دمایی و الگوهای بارش را درخود جای داده است. نیمی از وسعت آن را مناطق کوهستانی و نیمی دیگر را مناطق دشتی و کم ارتفاع تشکیل داده‌اند، مرتفع‌ترین نقطه آن در قله بینالود با ارتفاع ۳۳۳۹ متر و پست‌ترین نقطه آن در دشت سرخس با ارتفاع ۲۹۹ متر از سطح دریا واقع شده‌اند، اقلیم این استان براساس معیار طبقه‌بندی آمبرژه، اقلیم خشک سرد است و متوسط بارش سالانه آن حدود ۲۵۴ میلی‌متر درسال است. مناطق جنوبی این استان دارای بارشی کم‌تر از ۱۵۰ میلی‌متر در سال بوده و نواحی مرکزی و شمال غربی آن بیش از۳۲۰ میلی‌متر بارش در سال را دریافت می‌نمایند (رجبی و همکاران، ۱۴۰۲). شکل (۱) نقشه استان خراسان‌رضوی به همراه توزیع مکانی ایستگاه‌های باران‌سنج ثبات وزارت نیرو را نمایش می‌دهد.



شکل ۱- نقشه استان خراسان‌رضوی و پراکندگی مکانی ایستگاه‌های باران‌سنج ثبات وزارت نیرو

Figure 1- Map of Razavi Khorasan Province and spatial distribution of automatic rain gauge stations of the Ministry of Energy

داده‌های مورد استفاده

در این تحقیق از داده‌های بارش روزانه ۱۹ ایستگاه باران‌سنج ثبات وزارت نیرو که مشخصات، موقعیت و توزیع مکانی آن‌ها در جدول (۲) و شکل (۱) ارائه شده است، به همراه مقادیر بارش روزانه پایگاه داده ERA5 در محل ایستگاه‌های زمینی، برای سال‌های ۲۰۱۰ تا ۲۰۲۱ استفاده شده است.

[DOR: 20.1001.1.24235970.1403.12.1.8.1]

داده‌های ایستگاه‌های باران‌سنجی

برای دستیابی به بارش روزانه ایستگاه‌های باران سنج ثبات، اقدام به استخراج مقادیر بارش‌های رخداده با گام زمانی نیم ساعته با استفاده از نرم‌افزار RRDgitizer ver 2.1.1.1 از گراف آن‌ها شد. با توجه به اختلاف زمان موجود بین زمان تنظیمی باران سنج‌ها در منطقه مورد مطالعه با زمان مرجع بین‌المللی (UTC)[1]، زمان و میزان بارش‌ها با زمان مرجع منطبق شد و بارش روزانه با جمع مقادیر ساعتی محاسبه شد. این عمل برای کاهش خطای میزان تطابق بارش از دو مرجع متفاوت (ایستگاه‌های زمینی و داده‌های ERA5) و در دو سیستم زمانی مختلف صورت گرفت.

جدول ۲- مشخصات ایستگاه‌های باران‌سنجی ثبات وزارت نیرو مورد استفاده در پژوهش

Table 2 - Specifications of the automatic rain gauge stations of the Ministry of Energy used in the research

شناسه	ایستگاه	مشخصات جغرافیایی		
		UTM__X	UTM__Y	ارتفاع (m)
1	اداره مشهد	731051	4021779	970
2	ارداک بند ساروج	713648	4067524	1320
3	آل	738162	4066787	1452
4	باغسنگان تربت‌جام	281783	3900564	938
5	بلغور	731838	4030506	1920
6	تربت‌حیدریه	701468	3913151	1468
7	تلغور	710544	4078183	1540
8	چناران	690538	4057482	1170
9	درگز	685794	4145323	429
10	سد طرق	729885	4006132	1240
11	سد کارده	740695	4056093	1265
12	سرخس	334933	4044721	273
13	فریمان	757541	3953876	1415
14	کارخانه قند جوین	536166	4054892	1138
15	کارخانه قند نیشابور	643474	4017696	1199
16	کاشمر	632645	3899535	1054
17	گوش بالا	728919	4066333	1561
18	مغان	714070	4001739	1770
19	هی‌هی قوچان	636561	4108912	1333

پایگاه داده ERA5

ERA5 نسل پنجم تجزیه و بازتحلیل مقادیر پدیده‌های جوی است که داده‌های خروجی آن برگرفته از باز تحلیل اطلاعات استخراج شده از مدل ECMWF[2] است که تجزیه و باز تحلیل اطلاعات در آن با منظور نمودن تأثیر مقادیر شبکه ایستگاه‌های زمینی، انجام می‌شود، متغیرهای جوی در این پایگاه، دارای وضوح زمانی حداقل یک ساعته و دقت مکانی ۰/۲۵ × ۰/۲۵ درجه هستند و از سال ۱۹۷۵میلادی با تأخیر اندکی نسبت به زمان واقعی در دسترس قرار دارند (رجبی و همکاران، ۱۴۰۲). مقادیر بارش روزانه پایگاه داده ERA5، به لحاظ کمی و کیفی دارای داده‌هایی مناسب دراستان خراسان‌رضوی هستند (رجبی و همکاران، ۱۴۰۲)، از این رو از مقادیر بارش روزانه این پایگاه داده در فرایند تجزیه، تحلیل و واسنجی استفاده شد. مقادیر بارش روزانه پایگاه داده ERA5 از سامانه متن باز و تحت وب GEE[3] به آدرس ERA5_LAND/HOURLY استخراج شد.

روش تحقیق

برای ارزیابی کارایی الگوریتم‌های رگرسیونی یادگیری ماشین، برای محصول بارش روزانه ERA5 درمحل ایستگاه‌های باران‌سنج ثبات واقع در محدوده مورد مطالعه تحلیل‌های آماری انجام شد. مقادیر بارش روزانه ارائه شده توسط پایگاه داده ERA5، طی سال‌های ۲۰۱۰ تا ۲۰۲۱ (منطبق بر مهر ۱۳۸۹ تا پایان شهریور ۱۴۰۰) به مدت ۱۱ سال آبی، در پیکسل‌های متناظر با محل ۱۹ ایستگاه باران‌سنج ثبات واقع در محدوده مورد مطالعه استخراج شد و سپس با استفاده از الگوریتم‌های D-Tree، KNN، MLP اقدام به محاسبه مقادیر تصحیح شده بارش روزانه این پایگاه داده شد، پارامترهای آماری ضریب همبستگی (CC)، ریشه دوم میانگین مربع خطا (RMSE) و سوگیری (Bias) در محل هر ایستگاه، محاسبه و بهترین برآوردگر، جهت واسنجی وافزایش دقت، تصحیح سوگیری و کاهش خطای داده‌های بارش روزانه شناسایی شد. مقادیر بارش ایستگاه اداره مشهد به‌دلیل نامطلوب بودن داده‌های آن ازروند تجزیه و تحلیل‌ها کنار گذاشته شد.

[1] Coordinated Universal Time
[2] European Centre for Medium-Range Weather
[3] Google Earth Engine

[DOR: 20.1001.1.24235970.1403.12.1.8.1]

واسنجی داده‌های بارش حاصل از مدل با ایستگاه‌های شاهد، به‌وسیله الگوریتم‌های یادگیری ماشین به‌طور گسترده‌ای در حال استفاده و توسعه است، لذا تشخیص الگوی مناسب، نیازمند تحلیل خروجی داده‌های آزمایشی است، بنابراین تحلیل‌های آماری مقادیر آزمایشی ۱۸ ایستگاه زمینی و برای سه الگوریتم یاد شده به‌منظور بهینه‌سازی و واسنجی داده‌های بارش روزانه ERA5 صورت پذیرفت تا مشخص شود، کدام الگوریتم شاخص‌های خطا (Bias, RMSE) را کمینه و ضریب همبستگی (CC) را بیشینه می‌نماید و در کل، عملکرد بهتری ازخود به نمایش می‌گذارد.

پردازش داده‌ها

مقادیر بارش روزانه ایستگاه‌های باران‌سنج ثبات و داده‌های متناظر با آن‌ها در پایگاه داده ERA5 برای ۷۲،۲۷۰ روز بین سال‌های ۲۰۱۰ تا ۲۰۲۱ استخراج شد، با توجه به اختلاف زمان ثبت اطلاعات، بین باران‌سنج وزارت نیرو و این پایگاه داده (اختلاف زمان محلی و بین‌المللی)، مقادیربارش ایستگاه‌های زمینی از نظر زمانی بر داده‌های ERA5 منطبق شد و سپس با جمع مقادیر بارش کوتاه‌مدت، بارش روزانه برای هر دو منبع محاسبه شد. برای شناسایی و حذف داده‌های دورافتاده روش‌های مختلفی وجود دارد از جمله رسم نمودار جعبه‌ای با نمایه‌های چارکی، رسم نمودار پراکندگی داده‌ها، رسم نمودار Q-Q و هم‌چنین استفاده از مفاهیم آماری مانند تعریف یک حدآستانه، دراین پژوهش برای شناسایی داده‌های دورافتاده از حدآستانه استفاده شد، داده‌های بزرگ‌تر و کوچک‌تر از سه برابر انحراف معیار شناسایی و پس از بررسی و تطبیق آن با مقادیر مشابه در ایستگاه‌های مجاور در ارتباط با حذف و یا استفاده از آن‌ها تصمیم‌گیری شد. ۲۵ درصد داده‌ها به‌عنوان داده‌های آزمایشی و ۷۵ درصد باقی‌مانده برای آموزش مدل انتخاب شدند.

الگوریتم‌های یادگیری ماشین

در حوزه هوش مصنوعی و یادگیری ماشین، روش‌ها و الگوهای گوناگونی برای حل مسائل رگرسیونی وجود دارد که بسته به شرایط حاکم برپژوهش و نظر محقق از آن‌ها استفاده می‌شود. در این پژوهش ازسه الگوریتم KNN، D-Tree و MLP به‌منظور واسنجی مقادیر بارش روزانه پایگاه داده ERA5 استفاده شد.

رگرسیون K نزدیک‌ترین همسایگان (K-Nearest Neighbors)

الگوریتم K نزدیک‌ترین همسایگان یکی از محبوب ترین رویکردهای داده کاوی است و فرض اصلی آن این است که اگر همه K نزدیک‌ترین همسایه‌های یک نقطه در مجموعه آموزشی به یک دسته تعلق داشته باشند، فرض می‌شود که نقطه دارای ویژگی‌ها و کیفیت‌های یکسانی است. با استفاده از ویژگی‌های داده، KNN می‌تواند از روش حل معادله دشوار اجتناب نموده و بر همبستگی تمرکز نماید، الگوریتم KNN از شباهت ویژگی برای پیش‌بینی مقادیرهر نقطه داده جدید استفاده می‌کند. این بدان معنی است که نقطه جدید بر اساس شباهت بسیار نزدیک به نقاط مجموعه آموزشی است و بر همین اساس مقداری به آن اختصاص داده می‌شود. (فتاحی و جیریایی، ۱۴۰۰؛ Huang et al., 2017). اما آن‌چه در KNN اتفاق می‌افتد، به‌صورت گام به گام در زیر به آن اشاره شده است:

۱- ابتدا فاصله بین نقطه جدید و هر نقطه در داده‌های آموزشی محاسبه می‌شود. ۲- نزدیک‌ترین k نقطه داده بر اساس فاصله انتخاب خواهند شد. ۳- میانگین این نقاط داده، پیش‌بینی نهایی برای مقداردهی به نقطه جدید است.

اولین مرحله، محاسبه فاصله بین نقطه جدید و هر نقطه در داده‌های آموزشی است. روش‌های مختلفی برای محاسبه این فاصله وجود دارد که رایج‌ترین روش‌ها عبارتند از: اقلیدسی، منهتن و مینکوفسکی (برای داده‌های پیوسته)، فاصله مینکوفسکی شکل تعمیم یافته فاصله اقلیدسی و منهتن است.

- **فاصله اقلیدسی:** این فاصله به‌عنوان جذر مجذور تفاوت‌های یک نقطه جدید (x) و یک نقطه موجود (y) محاسبه می‌شود.

$$Euclidean = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \quad (۱)$$

- **فاصله منهتن:** این فاصله بین بردارهای واقعی و برابر مجموع اختلاف مطلق آن‌ها است.

$$Manhattan = \sum_{i=1}^{k}|x_i - y_i| \quad (۲)$$

[DOR: 20.1001.1.24235970.1403.12.1.8.1]

مقدار K، تعداد همسایه‌هایی را مشخص می‌کند که برای برآورد یک مشاهده جدید، به میانگین این تعداد همسایه نزدیک به آن مشاهده رجوع می‌شود. با توجه به تغییر در مقدار K، نتیجه نهایی براورد مقدارجدید، تمایل به تغییر پیدا می‌کند، لذا برای به‌دست آوردن مقدار بهینه K باید براساس محاسبه خطای داده‌های آموزشی و اعتبارسنجی آن تصمیم گیری شود. مقدار بهینه این ضریب، عددی است که درآن خطا در مقادیر آموزشی و آزمایشی به حداقل خود می‌رسد، زیرا به حداقل رسیدن عدم قطعیت هدف نهایی درحل مسائلی از این دست است.

رگرسیون درخت تصمیم (Decision Tree)

رگرسیون درخت تصمیم، یک الگوریتم یادگیری نظارت شده است که برای پیش‌بینی مقادیر پیوسته استفاده می‌شود. این الگوریتم با تقسیم فضای ویژگی، به مجموعه‌ای از مستطیل‌ها، و سپس اختصاص دادن مقدار ثابت به هر مستطیل عمل می‌نماید. بنابراین پیش‌بینی یک مشاهده جدید، برابر مقدار ثابت مستطیلی است که در آن قرار می‌گیرد. در واقع، الگوریتم رگرسیون درخت تصمیم با تقسیم بازگشتی فضای ویژگی به مناطق کوچک‌تر و برازش یک مدل ساده (معمولاً یک مقدار ثابت) برای هر منطقه کار می‌کند. این فرآیند توسط یک تابع ضرر کنترل می‌شود که خطای بین مقادیر پیش‌بینی شده و مقادیر واقعی در داده‌های آموزشی را اندازه‌گیری می‌کند (Chaudhary & Dhanya, 2020; Sandri & Zuccolotto, 2008). فرایند انجام رگرسیون درخت تصمیم را می‌توان به‌صورت زیر توصیف نمود:

- معیار تقسیم: در هر گره درخت تصمیم، این الگوریتم، یک ویژگی و یک مقدار آستانه را برای تقسیم داده‌ها به دو زیر مجموعه انتخاب می‌کند. این معیار تقسیم برای به حداقل رساندن تابع ضرراست، که با انتخاب یکی از شاخص‌های خطا، مانند میانگین مربعات خطا یا میانگین خطای مطلق، صورت می‌گیرد.
- پارتیشن‌بندی بازگشتی: الگوریتم درخت تصمیم، به تقسیم داده‌ها به زیر مجموعه‌های کوچک‌تر در هر گره، بر اساس معیار تقسیم انتخابی ادامه می‌دهد. این فرآیند تا زمانی ادامه می‌یابد که یک معیار توقف، مانند رسیدن به حداکثر عمق درخت یا حداقل داده در یک گره، تامین شود.
- برازش مدل: هنگامی که داده‌ها به مناطق کوچک‌تر تقسیم می‌شوند، یک مدل ساده (یا یک مقدار ثابت) برای هر منطقه برازش داده می‌شود. این مدل پیش‌بینی نقاط داده را در آن منطقه نشان می‌دهد.
- پیش‌بینی: برای پیش‌بینی یک نقطه داده جدید، الگوریتم، مسیر درخت تصمیم را طی می‌کند و مقدار ثابت ناحیه‌ای را که مشاهده جدید در آن قرار می‌گیرد را به‌عنوان مقدار پیش‌بینی شده اختصاص می‌دهد.

- پرسپترون‌های چند لایه (MLP)

پرسپترون چند لایه، یک شبکه عصبی مصنوعی با ساختار رو به جلو است که دارای لایه‌های مختلفی از گره‌ها است و هر کدام به لایه بعدی توسط همین گره‌ها متصل می‌شود. هر گره علاوه‌بر گره‌های ورودی از یک نورون با تابع فعال ساز غیرخطی تشکیل شده است. MLPها اغلب با استفاده از الگوریتم‌های پس انتشار آموزش داده می‌شوند و می‌توانند برای حل رگرسیون و طبقه‌بندی مسائل استفاده شوند. در ساختار یک MLP، گره‌های (نورون‌های) زیادی وجود دارد و در مجموع شامل سه لایه است. لایه ورودی، اولین لایه بوده و شامل متغیرهای مستقل برای پردازش است، لایه دوم در ساختار MLP، لایه‌های پنهان است که باید در میان لایه‌های خروجی و ورودی متمرکز شوند. راهبردهای‌های پیش‌خور، مقادیر را از یک لایه به لایه دیگر منتقل می‌کنند و بسته به نوع مسئله از یک یا چند لایه پنهان، بین ورودی و خروجی عبور نموده و از توابع انتقال برای تشخیص الگو موجود بین ورودی‌ها و خروجی‌ها استفاده می‌نمایند. MLP می‌تواند از انواع توابع انتقال استفاده کندکه رایج‌ترین آن‌ها "Sigmoid" و "Tansig" است. لایه سوم، که به لایه خروجی موسوم است، دربرگیرنده تخمین و یا پیش‌بینی مقادیر هدف است و در بیش‌تر موارد، "Pureline" به‌عنوان تابع انتقال لایه خروجی مورد استفاده قرار می‌گیرد. عملکرد یک شبکه MLP تحت تأثیر تعداد لایه‌ها و نورون‌های پنهان، در آن لایه‌ها است. آزمون و خطا یک روش معمول و آسان برای بهینه‌سازی این مقادیر است. گام بعدی پس از ترسیم ساختار MLP، انتخاب وزن‌ها و بایاس‌های متعدد برای هر نورون در بخش آموزش است، تا کارایی شبکه بهبود یابد. شبکه‌های MLP با استفاده از تکنیک‌های مختلفی آموزش داده می‌شود، از جمله منظم‌سازی بیزی (BR)، لونبرگ–مارکوارت (LM) و گرادیان مزدوج مقیاس شده (SCG) که این الگوریتم‌ها وظیفه محاسبه بایاس‌ها و وزن‌ها را بر عهده دارند. برحسب نیاز می‌توان از شبکه‌های عمیق و یا ساده استفاده نمود و این الگوریتم توانایی بسیار بالایی در یادگیری و آموزش دارد (قبائی سوق و همکاران، ۱۳۸۹؛ Khalili et al., 2016; Esteves et al., 2018).

توابع هزینه که می‌تواند در MLP رگرسیونی استفاده شوند عبارتند از میانگین مربعات خطا (MSE) و میانگین خطای مطلق (MAE)، هرچه مقدار خطا کم‌تر باشد مدل بهترین عملکرد را درآن محدوده ازخود به نمایش گذاشته است. این مدل رگرسیونی مربع

[DOR: 20.1001.1.24235970.1403.12.1.8.1]

خطاها را از طریق توابع LBFGS، Stochastic gradient descent و یا adam بهینه می‌نماید، آنچه درمدل رگرسیون پرسپترون چند لایه باید مدنظر قرار گیرد، تنظیم فراپارامترهای اجرای آن ازجمله، تعداد نورون‌ها، تابع فعال‌سازی برای لایه پنهان و تابع بهینه‌ساز، تعداد لایه پنهان و تعداد تکرار است که سعی و خطا درکمینه نمودن انحراف و به حداقل رساندن عدم قطعیت، ساده‌ترین روش برای به‌دست آوردن مقادیر بهینه است.

معیارهای ارزیابی

به‌طورکلی برای مقایسه بین چندین منبع تولید اطلاعات یکسان و تشخیص بهترین منبع از شاخص‌های آماری برای تعیین دقت محصولات استفاده می‌شود. در این پژوهش نیز از معیارهای آماری مختلفی برای ارزیابی و مقایسه داده‌های تخمین بارش ارائه شده توسط الگوریتم‌های یادگیری ماشین استفاده شد. توابع آماری و شاخص‌های مورد استفاده در جدول (۳) ارائه شده است. از ضریب همبستگی پیرسون (CC)[1]، برای نشان دادن میزان همبستگی بین مقادیر ثبت شده توسط باران‌سنج‌ها و مقادیر تخمینی توسط الگوریتم‌های یادگیری ماشین استفاده شد. همچنین، از میانگین ریشه دوم خطاها (RMSE)[2] برای ارزیابی مقدار خطا در مجموعه داده‌های تخمینی استفاده شد. از شاخص آماری سوگیری، برای ارزیابی تمایل الگوریتم‌ها، به بیش برآورد یا کم برآورد مقادیر بارش استفاده شد (Sorooshian et al., 1993).

جدول ۳- شاخص‌های آماری مورد استفاده در ارزیابی عملکرد الگوریتم‌های یادگیری ماشین درتخمین بارش روزانه

Table 3- Statistical indicators used in evaluating the performance of machine algorithms in estimating daily precipitation

رابطه	معادله	واحد	دامنه تغییرات	مقدار بهینه
1	$e_i = S_i - G_i$	متغیر	$(-\infty, +\infty)$	0
2	$CC = \frac{\frac{1}{n}\sum_{i=1}^{n}(S_i - \bar{S}) - (G_i - \bar{G})}{\sigma_S \sigma_G}$	بدون واحد	$[-1 \text{ و } 1]$	+1, -1
3	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2}$	متغییر	$(0, +\infty)$	0
4	$Bias = \frac{\sum_{i=1}^{n} e_i}{\sum_{i=1}^{n} G_i}$	متغییر	$(-\infty, +\infty)$	0

در روابط مندرج در جدول فوق، n تعداد نمونه، G مقادیر ثبت شده توسط باران‌سنج، S مقادیر شبیه‌سازی شده توسط الگوریتم، $\bar{S}$ و $\bar{G}$ به‌ترتیب میانگین مقادیر تخمین زده شده بارش توسط الگوریتم و باران‌سنج است. هم‌چنین تغییرات بارش توسط دو پارامتر σ_G و σ_S برآورد شد که نشان‌دهنده انحراف استاندارد مقادیر ثبت شده باران توسط باران‌سنج و الگوریتم است.

نتایج

ارزیابی شاخص‌های آماری

به‌منظور شناسایی بهترین الگوریتم در بهینه نمودن و کاهش انحراف مقادیر بارش روزانه پایگاه داده ERA5، اقدام به مقایسه، ارزیابی و تحلیل پارامترهای آماری سه الگوریتم یادگیری ماشینی مورد استفاده در این پژوهش شد.

ضریب همبستگی (CC)

ضریب همبستگی معیاری برای تشخیص درجه وابستگی داده‌های مدل با مقادیر مرجع است و دامنه تغییرات آن از منفی یک تا مثبت یک متغیر است و هرچه به کران‌های بالا و پایین نزدیک‌تر باشد بیانگر شدت وابستگی است و مقدارصفراین شاخص حاکی از عدم وجود ارتباط معنی‌دار بین دو دسته داده است، برای بررسی وضعیت ارتباط بین مقادیر بارش اصلاح شده ERA5 توسط الگوریتم‌های MLP، KNN و D-

[1] Correlation Coefficient
[2] Root Mean Square Error

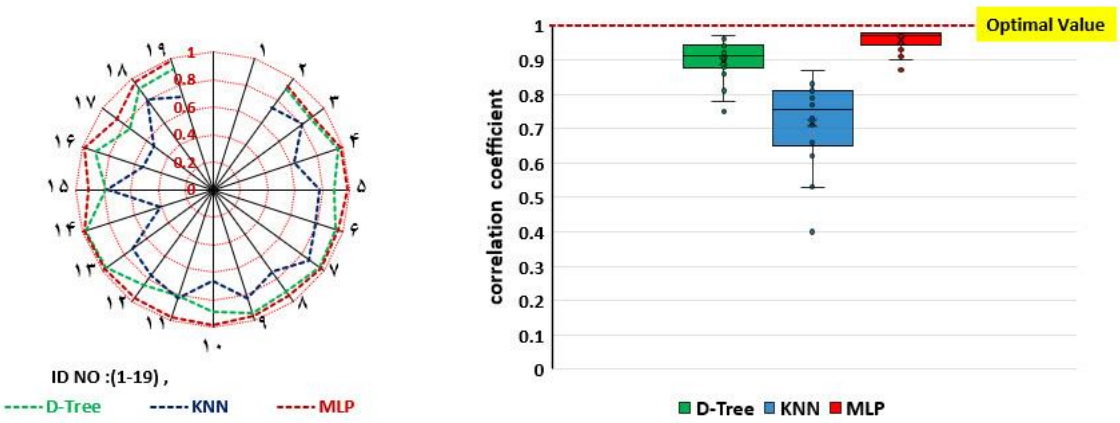
[DOR: 20.1001.1.24235970.1403.12.1.8.1]

Tree با بارش‌های ثبت شده در ایستگاه‌های زمینی، اقدام به محاسبه این ضریب برای هر ایستگاه و به تفکیک هر الگوریتم شد که نتایج آن در جدول (۴) و شکل (۲) ارائه شده است. دامنه تغییرات ضریب همبستگی در MLP بین ۰/۸۷ تا ۰/۹۸ متغیر بوده و این تغییرات برای D-Tree و KNN به‌ترتیب بین ۰/۷۵ تا ۰/۹۷ و ۰/۴ تا ۰/۸۷ محاسبه شده‌اند، در ۷۵ درصد ایستگاه‌ها ضریب همبستگی بین داده‌های اصلاحی و مقادیر مشاهده شده، در الگوریتم‌های MLP، D-Tree وKNN به‌ترتیب بیش‌تر از ۹۵، ۸۸ و ۶۵ درصد است. لذا، ضریب همبستگی درکلیه ایستگاه‌های مرجع با خروجی الگوریتم MLP ارتباط قوی‌تری داشته و از این حیث دارای عملکرد مناسب‌تری است.

جدول ٤- ضریب همبستگی مقادیرداده‌های آزمایشی بارش روزانه الگوریتم‌های مورد استفاده با ایستگاه‌های شاهد

Table 4- Correlation coefficient of experimental daily precipitation data values of used algorithms with control stations

ردیف	ایستگاه	ضر یب همبستگی			ردیف	ایستگاه	ضریب همبستگی		
		D-Tree	KNN	MLP			D-Tree	KNN	MLP
1	ارداک	0.91	0.73	0.93	10	سدکارده	0.81	0.83	0.91
2	آل	0.88	0.81	0.9	11	سرخس	0.86	0.77	0.97
3	باغسنگان	0.96	0.62	0.98	12	فریمان	0.96	0.73	0.98
4	بلغور	0.89	0.78	0.98	13	جوین	0.97	0.40	0.98
5	تربت‌حیدریه	0.94	0.79	0.96	14	نیشابور	0.78	0.77	0.91
6	تلغور	0.96	0.87	0.97	15	کاشمر	0.9	0.53	0.98
7	چناران	0.91	0.74	0.95	16	گوش	0.75	0.53	0.87
8	درگز	0.94	0.83	0.96	17	مغان	0.91	0.81	0.96
9	سدطرق	0.84	0.66	0.96	18	هی‌هی	0.92	0.71	0.96



شکل ۲- مقایسه شاخص آماری ضریب همبستگی بارش تخمینی داده‌های آزمایشی و مقادیر ایستگاهی با استفاده از الگوریتم‌های مورد استفاده

Figure 2- Comparison of the statistical index of the correlation coefficient of the estimated precipitation of experimental data and station values using the used algorithms

ریشه میانگین مربع خطاها (RMSE)

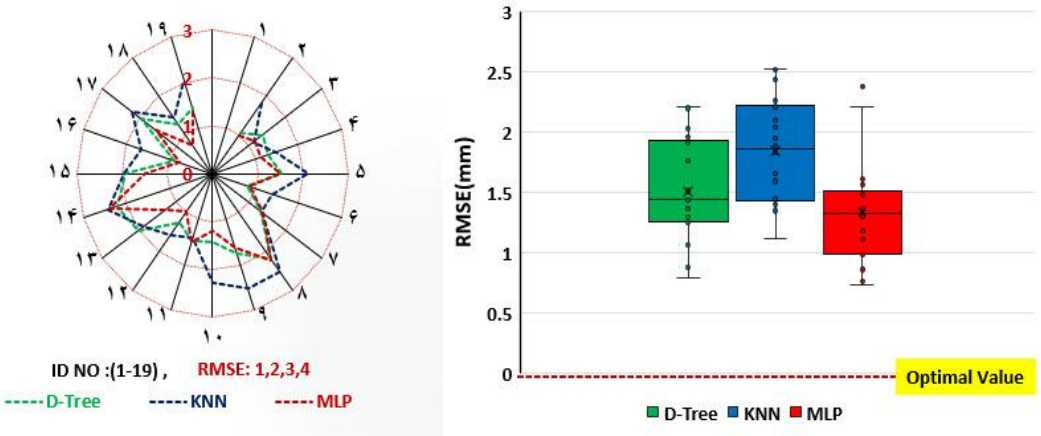
ریشه میانگین مربع خطا، معیاری برای تشخیص عملکرد مدل در تولید داده است، دامنه تغییرات این شاخص بین صفر تا بی‌نهایت بوده و هرچه به صفر نزدیک‌تر باشد، نشان از عملکرد مناسب مدل در تخمین مقادیر بارش است. این شاخص، دارای مقیاس و واحد مشابه با داده‌های مورد تحلیل است و همواره مثبت و بدون جهت است. به‌منظور بررسی وضعیت تغییرات این پارامتر و عملکرد الگوریتم‌های مورد استفاده، اقدام به محاسبه، مقایسه و تحلیل این شاخص برای هرایستگاه و به تفکیک هر الگوریتم (MLP,KNN,D-Tree) شد، نتایج محاسبه این پارامتر آماری درجدول (۵) و شکل (۳) ارائه شده است. دامنه تغییرات RMSE درMLP بین ۰/۷ تا ۲/۴ میلی‌متر در روز متغیر بوده و این تغییرات برای D-Tree و KNN به‌ترتیب بین ۰/۸ تا ۲/۲ و ۱/۲ تا ۲/۵ محاسبه شده‌اند، در ۷۵ درصد ایستگاه‌ها RMSE داده‌های اصلاحی و مقادیر مشاهده شده، در الگوریتم‌های MLP، D-Tree و KNN به‌ترتیب کم‌تر از ۱/۵، ۱/۹ و ۲/۲ میلی‌متر در روز است. RMSE در اغلب ایستگاه‌های مرجع، درالگوریتم MLP به مقدار بهینه، نزدیک‌تر از KNN و D-Tree بوده لذا دارای عملکرد مناسب‌تری در مقایسه با دو الگوریتم دیگر است.

[DOR: 20.1001.1.24235970.1403.12.1.8.1]

جدول ۵- مقادیر RMSE داده‌های آزمایشی بارش روزانه الگوریتم‌های مورد استفاده با ایستگاه‌های شاهد

Table 5- RMSE of daily test data of algorithms used with reference stations

ردیف	ایستگاه	ریشه میانگین مربع خطا-میلی‌متر			ردیف	ایستگاه	ریشه میانگین مربع خطا-میلی‌متر			
		D-Tree	KNN	MLP			D-Tree	KNN	MLP	
1	ارداک	1.1	1.8	1.0	10	سدکارده	1.5	1.4	1.4	1.5
2	آل	1.4	1.1	1.2	11	سرخس	1.2	1.6	1.6	1
3	باغسنگان	1.3	1.5	1.1	12	فریمان	2.0	1.9	1.9	1.3
4	بلغور	1.5	2.1	1.5	13	جوین	2.2	2.4	2.4	2.4
5	تربت‌حیدریه	0.8	1.4	0.9	14	نیشابور	1.9	2.0	2.0	1.5
6	تلغور	1.3	1.3	1.4	15	کاشمر	0.9	1.7	1.7	0.8
7	چناران	2.2	2.5	2.2	16	گوش	2.0	2.2	2.2	1.6
8	درگز	1.8	2.5	1.6	17	مغان	1.3	1.4	1.4	0.7
9	سدطرق	1.4	2.3	1.2	18	هی‌هی	1.4	2.0	2.0	1.4



شکل ۳- مقایسه شاخص آماری RMSE بارش تخمینی داده‌های آزمایشی ومقادیرایستگاهی با استفاده ازالگوریتم‌های مورد استفاده

Figure 3- Comparison of RMSE statistical index of estimated precipitation of experimental data and station values using the used algorithms

سوگیری

این شاخص معیاری برای تشخیص سوگیری داده‌های تولید شده توسط مدل است که دارای مقدار و جهت است، دامنه تغییرات این شاخص محدودیتی نداشته و مقدار بهینه آن صفر بوده و دیمانسیون این پارامتر آماری با داده‌های مورد تحلیل یکسان است. مقدار عددی سوگیری میزان انحراف داده‌های اصلاحی از مقادیر حقیقی را ارائه می‌دهد و علامت آن تمایل الگوریتم به بیش برآورد و یا کم برآورد نسبت به مقادیر مشاهداتی را نشان می‌دهد. به‌منظور بررسی وضعیت تغییرات این پارامتر، جهت‌گیری آن و عملکرد الگوریتم‌های مورد استفاده، اقدام به محاسبه، مقایسه و تحلیل این شاخص برای کلیه ایستگاه‌های مرجع و به تفکیک الگوریتم‌های مورد استفاده، شد، نتایج محاسبه این پارامتر آماری در جدول (۶) و شکل (۴) ارائه شده است. دامنه تغییرات سوگیری درMLP، [-0.6,0.18] میلی‌متر در روز بوده و این دامنه تغییرات برای D-Tree و KNN به‌ترتیب [۰/۱۶، ۰/۵-] و [۰/۶، ۰/۸-] محاسبه شده‌اند، سوگیری داده‌های اصلاحی و مقادیر مشاهده شده، در الگوریتم‌های MLP، D-Tree وKNN برای میانه ایستگاه‌ها به‌ترتیب برابر ۰/۰۹-، ۰/۱۱- و ۰/۱۶- میلی‌متر در روز است. سوگیری مقادیر اصلاح شده در اکثر ایستگاه‌ها دارای جهت‌گیری منفی بوده و نشان ازتمایل الگوریتم به کم برآورد داده‌های تخمینی نسبت به مقادیر واقعی دارد. با توجه به مقادیرسوگیری درسه الگوریتم و تراکم این مقادیر حول مقدار بهینه در MLP ، لذا پرسپترون چندلایه دارای عملکرد مناسب‌تری در مقایسه با دو الگوریتم دیگر در این زمینه است.

عملکرد الگوریتم منتخب

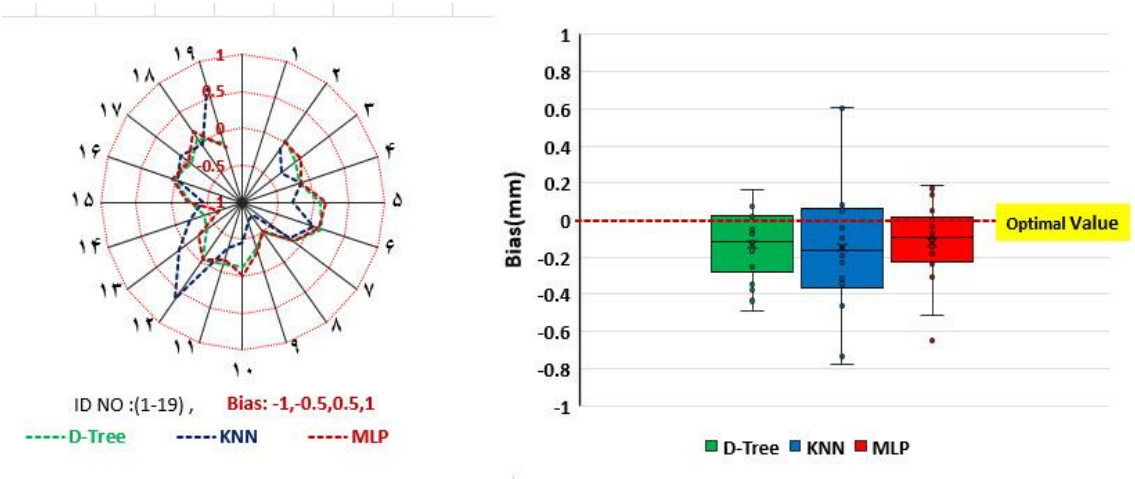
[DOR: 20.1001.1.24235970.1403.12.1.8.1]

با توجه به تحلیل‌های صورت گرفته روی داده‌های آزمایشی الگوریتم‌های مورد استفاده و مقادیر بارش روزانه ایستگاه‌های زمینی، مدل MLP دارای عملکرد بهتری نسبت به دو الگوریتم دیگر است، بنابراین به‌عنوان الگوی منتخب در واسنجی داده‌های بارش روزانه ERA5 در محدوده استان خراسان‌رضوی مناسب تشخیص داده شد. اما مقادیر بارش روزانه پایگاه داده ERA5 نسبت به مقادیر ایستگاه‌های زمینی دارای ضریب همبستگی، ریشه میانگین مربع خطا و سوگیری است، لذا باید توانایی الگوریتم منتخب در بهبود وضعیت داده‌ها وارتقاء شاخص‌های آماری مورد بررسی قرارگیرد. برای این منظور نسبت به ارزیابی، مقایسه و تحلیل شاخص‌های CC، RMSE و Bias مقادیر بارش روزانه پایگاه داده ERA5 و مقادیر واسنجی شده بارش توسط الگوریتم MLP برای ایستگاه‌های بارانسنجی اقدام شد.

جدول ٦- مقادیر سوگیری داده‌های آزمایشی بارش روزانه الگوریتم‌های مورد استفاده با ایستگاه‌های شاهد

Table 6- Bias values of experimental daily precipitation data of algorithms used with reference stations

ردیف	ایستگاه	سوگیری- میلی‌متر در روز			ردیف	ایستگاه	سوگیری- میلی‌متر در روز		
		D-Tree	KNN	MLP			D-Tree	KNN	MLP
1	ارداک	-0.037	-0.09	0.008	10	سدکارده	-0.123	-0.343	-0.18
2	آل	-0.078	-0.31	0.003	11	سرخس	-0.051	0.061	-0.03
3	باغسنگان	-0.147	-0.14	-0.14	12	فریمان	-0.378	0.099	-0.23
4	بلغور	-0.091	-0.28	0.171	13	جوین	-0.435	-0.232	-0.64
5	تربت‌حیدریه	0.162	0.054	0.137	14	نیشابور	-0.254	-0.445	-0.21
6	تلغور	-0.109	-0.18	-0.11	15	کاشمر	-0.021	-0.036	-0.05
7	چناران	-0.488	-0.78	-0.52	16	گوش	-0.104	0.083	-0.07
8	درگز	-0.343	-0.73	-0.31	17	مغان	0.075	-0.041	0.18
9	سدطرق	-0.122	-0.46	-0.01	18	هی‌هی	-0.163	0.603	-0.22

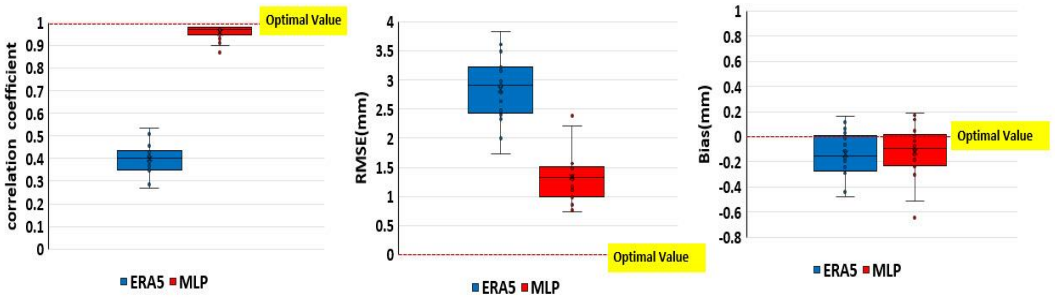


شکل ٤- مقایسه شاخص آماری سوگیری بارش تخمینی داده‌های آزمایشی و مقادیر ایستگاهی با استفاده از الگوریتم‌های مورد استفاده

Figure 4- Comparison of the statistical bias index of the estimated precipitation of experimental data and station values using the used algorithms

با توجه به شاخص‌های آماری به‌دست آمده، دامنه تغییرات ضریب همبستگی مقادیر واسنجی شده بین ۰/۸۷ تا ۰/۹۸ متغیر بوده درصورتی که تغییرات این شاخص برای داده‌های ERA5 بین ۰/۲۷ تا ۰/۵۳ استکه نشان‌دهنده بهبود داده‌های واسنجی شده توسط الگوریتم منتخب است. هم‌چنین دامنه تغییرات ریشه میانگین مربع خطاها ی داده‌های واسنجی شده بارش روزانه بین ۰/۷۵ تا ۲/۳۸ میلی‌متر در روز است در صورتی‌که این دامنه برای مقادیر ERA5 بین ۱/۷ تا ۳/۸ است که نشان‌دهنده بهبود وضعیت درمقادیر واسنجی شده است. مقدار سوگیری بارش روزانه واسنجی شده توسط الگوریتم منتخب برای میانه ایستگاه‌ها 0.09- میلی‌متر بوده و همین شاخص، برای مقادیر متناظر ERA5 ۰/۱۶- است که نشان‌دهنده عملکرد مناسب الگوریتم منتخب در واسنجی داده‌ها است.

[DOR: 20.1001.1.24235970.1403.12.1.8.1]



شکل ۵- مقایسه شاخص‌های آماری بارش روزانه ERA5 واسنجی شده برای ایستگاه‌های شاهد

Figure 5- Comparison of ERA5 daily precipitation statistical indices calibrated for control stations

بحث

مقادیر ارائه شده توسط سنجنده‌ها و پایگاه‌های داده، با کاهش مقیاس زمانی از داده‌های مشاهداتی فاصله بیش‌تری می‌گیرند، لذا برای استفاده از آن‌ها ضروریست تا این مقادیر با استفاده از روش‌ها و الگوهای پرکاربرد اصلاح شوند. این پژوهش نشان داد که در مقیاس روزانه، الگوریتم‌های یادگیری ماشین می‌توانند کمیت داده‌های بارش ERA5 را بهبود بخشیده و شاخص‌های آماری خطا را کاهش دهند و بارش را تا حد زیادی به مقادیر واقعی نزدیک نمایند. هر چند این الگوریتم‌ها قادربه بهبود شرایط هستند، اما عملکرد آن‌ها یکسان نبوده و نتایج متفاوتی را ارائه می‌نمایند، لذا ایجاب می‌نماید تا بهترین الگوریتم درواسنجی داده‌ها شناسایی شود. ارزیابی و تحلیل‌های آماری صورت گرفته مبین این است که الگوریتم MLP نسبت به D-Tree و KNN از عملکرد بهتری برخوردار بوده و می‌تواند شاخص‌های CC، RMSE و Bias را به‌طور موثری بهبود بخشد. هم‌چنین ارزیابی و تحلیل پارامترهای آماری برای مقادیر بارش روزانه اصلاح شده توسط این الگوریتم در مقایسه با داده‌های اولیه ERA5 نشان داد که MLP در ارتقاء ضریب همبستگی و کاهش RMSE و سوگیری بسیار خوب عمل نموده و توانایی بهبود داده‌های ERA5 را دارد.

با توسعه هوش مصنوعی و جنبه‌های کاربردی آن در حوزه‌ها و علوم مختلف، امروزه شاهد استفاده از یادگیری ماشین درعلوم آب و هواشناسی هستیم، لذا در سطح دنیا مطالعاتی در خصوص قابلیت استفاده از الگوریتم‌ها و الگوهای مختلف برای اصلاح سوگیری و کاهش خطا ارائه شده است، هرچند روش‌های قدیمی قادر به اصلاح داده‌ها هستند، اما فرایند انجام و تحلیل، توسط آن‌ها پیچیده و زمان بر است، تکنیک‌های یادگیری ماشین با توجه به سرعت، دقت، سادگی در قابلیت تکرار و ارائه نتایج خیره‌کننده، مورد اقبال قرار گرفته به‌طوری که استفاده از آن‌ها با سرعت درحال فراگیرشدن است. نتایج این پژوهش، همانند اغلب تحقیقات صورت گرفته استفاده از الگوریتم‌های یادگیری ماشین را برای بهبود عملکرد و کاهش خطا و واسنجی داده‌های بارش توصیه می‌نماید.

برای کاهش سوگیری داده‌های بارش سنجنده GPM Chaudhary & (۲۰۲۰) طی مطالعه‌ای اقدام به واسنجی مقادیربارش ماهانه این سنجنده نسبت به داده‌های شبکه‌ای ایستگاه‌های زمینی در شبهه قاره هند نمود، اوبا تحلیل داده‌های بارش ماهانه بین سال‌های ۲۰۰۱ تا ۲۰۱۶ میلادی واستفاده از روش درخت تصمیم، سوگیری آن‌ها را تصحیح نمود و نشان داد این الگوریتم قادر است به‌طور متوسط ضریب همبستگی را به ۰/۷۷ رسانده و شاخص RMSE را بین ۸/۷ تا ۳۷/۳ و MAB را بین ۶/۳ تا ۲۹/۲ بهبود ببخشد. این پژوهش نیز نشان داد که الگوریتم درخت تصمیم درافزایش ضریب همبستگی و کاهش شاخص‌های خطا موثر بوده و می‌تواند در بهبود شرایط نقش بسزایی ایفا نماید (Chaudhary & Dhanyav, 2020)، هم‌چنین مینگ مینگ (۲۰۱۷)، با استفاده از الگوریتم KNN و توسعه آن نشان داد این روش از عملکرد مناسبی در اصلاح و پیش‌بینی بارش در حوضه‌های آبریز کشور چین برخوردار است (Huang, 2017)، نتایج به‌دست آمده از این تحقیق نیز نشان داد که الگوریتم KNN قادر است شاخص‌های آماری‌خطا را کاهش داده و همبستگی را بهبود بخشد. خلیلی و همکاران (۲۰۱۶) از الگوریتم ANN برای پیش‌بینی بارش ماهانه استفاده نمود، او از آمار ماهانه ایستگاه سینوپتیک مشهد طی سال‌های ۱۹۵۳ تا ۲۰۰۳ و یک شبکه پرسپترون سه لایه با ریتم الگوی انتشار برگشتی بهره جست و با تنظیم فراپارامترهای مدل برای بارش ماهانه نشان داد که ضریب همبستگی ۰/۹۳ و مقادیر خطای RMSE وMAE به‌ترتیب برابر ۰/۹۹ و ۶/۰۲ است که با مقایسه با روش سنتی و رایج در سازمان هواشناسی (پیش‌بینی از روی نقشه‌های سینوپتیک) دارای دقت قابل قبولی در پیش‌بینی بارش ماهانه است (Khalili, et al, 2016)، خروجی‌های این مطالعه نیز نشان داد که الگوریتم MLP توانایی اصلاح و بهبود داده‌های اولیه را به‌طور مطلوبی دارد و می‌توان از آن در اصلاح سوگیری استفاده نمود. در مطالعه‌ای که به‌منظور پیش‌بینی بارش توسط الگوریتم پرسپترون چند لایه روی ده ناحیه متفاوت در کشور برزیل انجام شد، بارش و دمای ماهانه یک دوره ۶۰ ساله را به‌عنوان ورودی مدل جمع‌آوری شد و استفاده از الگوریتم MLP

[DOR: 20.1001.1.24235970.1403.12.1.8.1]

نشان داد که دقت مدل در فصل‌های مختلف متفاوت بوده و رابطه مستقیمی با ارتفاع و حجم نرمال بارش دارد (Esteves et al., 2018). Meyer و همکاران (۲۰۱۶) اقدام به اصلاح مقادیر بارش روزانه در کشور آلمان نمود و از میان الگوریتم‌های مورد استفاده (RF, ANN, SVR). جنگل تصادفی به‌عنوان بهترین الگوریتم تصحیح بارش تشخیص داده شد (Meyer et al., 2016)، در این پژوهش MLP به‌عنوان بهترین الگوریتم شناسایی شد و نتایج آن با RF وSVR مقایسه نشده است و در مطالعات تکمیلی باید به ارزیابی آن‌ها پرداخته شود. TAO و همکاران (۲۰۱۶) نیز کار مشابهی در مرکز آمریکا انجام و الگوهای یادگیری عمیق را مورد بررسی قرار دادند. آن‌ها پیشنهاد نمود که از شبکه عصبی پیچشی برای تصحیح بارش می‌توان به نحو مناسبی استفاده نمود. Baez و همکاران (۲۰۲۰) با پژوهشی در کشور شیلی، الگوریتم جنگل تصادفی (RF) را برای تصحیح سوگیری مناسب تشخیص دادند. Chen و همکاران (۲۰۲۰) با مطالعه‌ای در منطقه دالاس کشور آمریکا، الگوریتم‌های یادگیری عمیق را برای تصحیح بارش مناسب ارزیابی نمودند. Nguyen (۲۰۲۱) با مطالعه‌ای در کشورکره الگوریتم جنگل تصادفی را برای تصحیح بارش ماهواره‌ای پیشنهاد نمودند. این تحقیق نیز نتایج به‌دست آمده از تحقیقات سایر محققان در خصوص قابلیت الگوریتم‌های یادگیری ماشین درکاهش سوگیری را تایید می‌نماید.

نتیجه‌گیری

الگوریتم‌های یادگیری ماشین، قابلیت تصحیح مقادیر بارش از منابع دیگر از جمله ماهواره‌ها و پایگاه‌های داده را به‌خوبی دارا بوده و می‌توانند در مطالعات مورد استفاده قرارگیرند. نتایج به‌دست آمده از این پژوهش مبین آن است که استفاده از الگوریتم MLP برای تصحیح بارش‌های کوتاه‌مدت ماهواره‌ای در استان خراسان‌رضوی بسیارکارآمد بوده و از نتایج آن می‌توان در مطالعات مرتبط با علوم آب وهواشناسی استفاده نمود. تکنیک های یادگیری ماشین در تصحیح Bias، تنها نقطه ضعف سنجنده‌ها که عدم تطابق مقادیر آن‌ها با داده‌های مشاهداتی است را تا حد زیادی مرتفع نموده و نقطه عطفی در پردازش داده‌های ماهواره‌ای و استفاده گسترده از آن‌ها گشوده است.

در این پژوهش از داده‌های کوتاه‌مدت پایگاه ERA5 برای شناسایی بهترین الگوریتم از میان MLP، KNN، و D-Tree به‌منظور تصحیح بارش روزانه استفاده شد، لذا پیشنهاد می‌شود تا در تحقیقات دیگر:

- بارش‌های ماهانه و سالانه مورد تجزیه و تحلیل قرارگیرد.
- از سایر سنجنده‌ها و پایگاه‌های داده درمقیاس زمانی کوتاه‌مدت و بلندمدت استفاده شود.
- ازسایرالگوریتم‌های رگرسیونی مبتنی بر یادگیری ماشین برای تصحیح بارش استفاده به عمل آید.
- دقت تکنیک‌های یادگیری عمیق در تصحیح بارش، مورد آزمون قرارگیرد.
- الگوریتم‌های یادگیری ماشین در ریزمقیاس‌نمایی و تصحیح مقادیر بارش ماهواره‌ای و پایگاه داده بر سامانه GEE مورد بررسی و واسنجی قرارگیرد.

ملاحظات اخلاقی

دسترسی به داده‌ها: دسترسی به داده‌ها و نتایج استفاده شده در این پژوهش از طریق مکاتبه با نویسنده مسئول امکان‌پذیر است.

حمایت مالی: این پژوهش مستخرج از رساله دکتری بوده و تحت حمایت دانشگاه فردوسی مشهد به انجام رسیده است.

مشارکت نویسندگان: این پژوهش توسط مجید رجبی جاغرق بصورت طرح اولیه تهیه شد. سپس توسط آقایان دکتر سیدمحمد موسوی بایگی، دکتر سیدعلیرضا عراقی و دکترهادی جباری نوقابی اصلاح و تکمیل شد.

تضاد منافع نویسندگان: نویسندگان این مقاله اعلام می‌دارند که هیچ‌گونه تضاد منافعی در خصوص نگارش این پژوهش ندارند.

سپاس‌گزاری: از شرکت آب منطقه‌ای خراسان رضوی به‌دلیل همکاری صمیمانه در ارائه آمار و اطلاعات مورد نیاز تشکر و قدردانی می‌شود.

منابع

۱. رجبی جاغرق، مجید، موسوی بایگی، سید محمد، عراقی، سید علی رضا و جباری نوقابی، هادی (۱۴۰۲). ارزیابی دقت مقادیر بارش روزانه TRMM، GPM، ERA5 و PERSIANN در استان خراسان رضوی. *سامانه‌های سطوح آبگیر باران*، ۱۱(۲)، ۷۹-۱۰۱. **http://jircsa.ir/article-1-505-fa.html**

۲. فتاحی، هادی و جیریایی، فاطمه (۱۴۰۰). ارزیابی پتانسیل روانگرایی خاک در اثر وقوع زمین لرزه بااستفاده از چند الگوریتم طبقه‌بندی هوشمند در نرم‌افزار Orange، *عمران فردوسی*، ۳۴(۳)، ۳۹-۵۲. **https://dorl.net/dor/20.1001.1.27832805.1400.34.3.3.9**

۳. قبائی سوق، محمد، مساعدی، ابوالفضل، حسام، موسی و هزارجریبی، ابوطالب (۱۳۸۹). ارزیابی تأثیر پیش‌پردازش پارامترهای ورودی به شبکه عصبی مصنوعی (ANNs) با استفاده از روش‌های رگرسیون گام بهگام و گاما تست به‌منظور تخمین سریع‌تر تبخیر و تعرق روزانه، *آب و خاک*، ۲۴(۳)، ۶۱۰-۶۲۴. **doi.org/10.22067/jsw.v0i0.3631**

References

1. Abdollahipour, A., Ahmadi, H., & Aminnejad, B. (2022). A review of downscaling methods of satellite-based precipitation estimates. *Earth Science Informatics*, *15*(1), 1-20. doi.org/10.1007/s12145-021-00669-4
2. Amengual, A., Homar, V., Romero, R., Alonso, S., & Ramis, C. (2012). A statistical adjustment of regional climate model outputs to local scales: application to Platja de Palma, Spain. *Journal of Climate*, *25*(3), 939-957. doi.org/10.1175/jcli-d-10-05024.1
3. Baez-Villanueva, O. M., Zambrano-Bigiarini, M., Beck, H. E., McNamara, I., Ribbe, L., Nauditt, A., & Thinh, N. X. (2020). RF-MEP: A novel Random Forest method for merging gridded precipitation products and ground-based measurements. *Remote Sensing of Environment*, *239*, 111606. doi.org/10.1016/j.rse.2019.111606
4. Beck, H. E., Vergopolan, N., Pan, M., Levizzani, V., Van Dijk, A. I., Weedon, G. P., & Wood, E. F. (2017). Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling. *Hydrology and Earth System Sciences*, *21*(12), 6201-6217. doi.org/10.5194/hess-21-6201-2017
5. Blöschl, G., Bierkens, M. F., Chambel, A., Cudennec, C., Destouni, G., Fiori, A., & Renner, M. (2019). Twenty-three unsolved problems in hydrology (UPH)–a community perspective. *Hydrological Sciences Journal*, *64*(10), 1141-1158. doi.org/10.1080/02626667.2019.1620507
6. Chaudhary, S., & Dhanya, C. T. (2020). Decision tree-based reduction of bias in monthly IMERG satellite precipitation dataset over India. *h2oj*, *3*(1), 236-255. doi.org/10.2166/h2oj.2020.124
7. Chen, C., Hu, B., & Li, Y. (2021). Easy-to-use spatial Random Forest-based downscaling-calibration method for producing high resolution and accurate precipitation data. *Hydrology and Earth System Sciences Discussions*, *2021*, 1-50. 25, doi.org/10.5194/hess-25-5667-2021
8. Chen, H., Chandrasekar, V., Cifelli, R., & Xie, P. (2019). A machine learning system for precipitation estimation using satellite and ground radar network observations. *IEEE Transactions on Geoscience and Remote Sensing*, *58*(2), 982-994. doi.org/10.1109/tgrs.2019.2942280
9. Chen, H., Sun, L., Cifelli, R., & Xie, P. (2021). Deep learning for bias correction of satellite retrievals of orographic precipitation. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1-11. doi.org/10.1109/tgrs.2021.3105438
10. Chen, J., Brissette, F. P., Chaumont, D., & Braun, M. (2013). Finding appropriate bias correction methods in downscaling precipitation for hydrologic impact studies over North America. *Water Resources Research*, *49*(7), 4187-4205. doi.org/10.1002/wrcr.20331
11. Chen, S., Xiong, L., Ma, Q., Kim, J. S., Chen, J., & Xu, C. Y. (2020). Improving daily spatial precipitation estimates by merging gauge observation with multiple satellite-based precipitation products based on the geographically weighted ridge regression method. *Journal of Hydrology*, *589*, 125156. doi.org/10.1016/j.jhydrol.2020.125156
12. Curtis, S., Crawford, T. W., & Lecce, S. A. (2007). A comparison of TRMM to other basin-scale estimates of rainfall during the 1999 Hurricane Floyd flood. *Natural Hazards*, *43*, 187-198. http://dx.doi.org/10.1007/s11069-006-9093-y
13. Esteves, J. T., de Souza Rolim, G., & Ferraudo, A. S. (2019). Rainfall prediction methodology with binary multilayer perceptron neural networks. *Climate Dynamics*, *52*, 2319-2331. doi.org/10.1007/s00382-018-4252-x
14. Fattahi, H., & Jiryaee, F. (2021). Evaluation of Soil Liquefaction Potential due to Earthquake using Intelligent Classification Algorithm in Orange Software. *Journal of Ferdowsi Civil Engineering*, *34*(3), 39-59. https://dorl.net/dor/20.1001.1.27832805.1400.34.3.3.9 [in Persian].
15. Fernandez-Palomino, C. A., Hattermann, F. F., Krysanova, V.; Lobanova, A., Vega-Jácome, F., Lavado, W., ... & Bronstert, A. (2022). A novel high-resolution gridded precipitation dataset for Peruvian and Ecuadorian watersheds: Development and hydrological evaluation. *Journal of Hydrometeorology*, *23*(3), 309-336. doi.org/10.2139/ssrn.4602668
16. Ghabaei Sough, M., Mosaedi, A., Hesam, M. O. U. S. A., & Hezarjaribi, A. (2010). Evaluation effect of input parameters preprocessing in artificial neural networks (Anns) by using stepwise regression and gamma test techniques for fast estimation of daily evapotranspiration. *Water and Soil*, 24(3), 610-624. doi.org/10.22067/jsw.v0i0.3631[in Persian].
17. Gómez-Chova, L., Tuia, D., Moser, G., & Camps-Valls, G. (2015). Multimodal classification of remote sensing images: A review and future directions. *Proceedings of the IEEE*, *103*(9), 1560-1584. doi.org/10.1109/jproc.2015.2449668
18. Guo RuiFang, G. R., Liu YuanBo, L. Y., Zhou Han, Z. H., & Zhu YaQiao, Z. Y. (2018). Precipitation downscaling using a probability-matching approach and geostationary infrared data: an evaluation over six climate regions. *Hydrology and Earth System Sciences*, 22(7), 3685–3699. doi.org/10.5194/hess-22-3685-2018

19. Hamill, T. M., & Scheuerer, M. (2018). Probabilistic precipitation forecast postprocessing using quantile mapping and rank-weighted best-member dressing. *Monthly Weather Review*, 146(12), 4079-4098. doi.org/10.1175/mwr-d-18-0147.1
20. He, X., Chaney, N. W., Schleiss, M., & Sheffield, J. (2016). Spatial downscaling of precipitation using adaptable random forests. *Water Resources Research*, *52*(10), 8217-8237. doi.org/10.1002/2016wr019034
21. Heredia, M. B., Junquas, C., Prieur, C., & Condom, T. (2018). New statistical methods for precipitation bias correction applied to WRF model simulations in the Antisana region, Ecuador. *Journal of Hydrometeorology*, *19*(12), 2021-2040. doi.org/10.1175/jhm-d-18-0032.1
22. Hu, Q., Li, Z., Wang, L., Huang, Y., Wang, Y., & Li, L. (2019). Rainfall spatial estimations: A review from spatial interpolation to multi-source data merging. *Water*, *11*(3), 579. doi.org/10.3390/w11030579
23. Huang, M., Lin, R., Huang, S., & Xing, T. (2017). A novel approach for precipitation forecast via improved K-nearest neighbor algorithm. *Advanced Engineering Informatics*, *33*, 89-95. doi.org/10.1016/j.aei.2017.05.003
24. Jakob Themeßl, M., Gobiet, A., & Leuprecht, A. (2011). Empirical-statistical downscaling and error correction of daily precipitation from regional climate models. *International Journal of Climatology*, *31*(10), 1530-1544.doi.org/10.1002/joc.2168
25. Khalili, N., Khodashenas, S. R., Davary, K., Baygi, M. M., & Karimaldini, F. (2016). Prediction of rainfall using artificial neural networks for synoptic station of Mashhad: a case study. *Arabian Journal of Geosciences*, *9*, 1-9. doi.org/10.1007/s12517-016-2633-1
26. Li, C. Y. Climate Dynamics, 2nd ed.; Chapter 1; Meteorological Press: Beijing, China, 2000; pp. 2–3.
27. Li, H., Haugen, J. E., & Xu, C. Y. (2018). Precipitation pattern in the Western Himalayas revealed by four datasets. *Hydrology and Earth System Sciences*, *22*(10), 5097-5110. doi.org/10.5194/hess-22-5097-2018.
28. Lin, Q., Peng, T., Wu, Z., Guo, J., Chang, W., & Xu, Z. (2022). Performance evaluation, error decomposition and Tree-based Machine Learning error correction of GPM IMERG and TRMM 3B42 products in the Three Gorges Reservoir Area. *Atmospheric Research*, *268*, 105988. doi.org/10.1016/j.atmosres.2021.105988
29. Liu, X., Yang, T., Hsu, K., Liu, C., & Sorooshian, S. (2017). Evaluating the streamflow simulation capability of PERSIANN-CDR daily rainfall products in two river basins on the Tibetan Plateau. *Hydrology and Earth System Sciences*, *21*(1), 169-181. doi.org/10.5194/hess-21-169-2017
30. Ma, Y., Zhang, Y., Yang, D., & Farhan, S. B. (2015). Precipitation bias variability versus various gauges under different climatic conditions over the Third Pole Environment (TPE) region. *International Journal of Climatology*, *35*(7), 1201-1211. doi.org/10.1002/joc.4045
31. Mega, T., Ushio, T., Takahiro, M., Kubota, T., Kachi, M., & Oki, R. (2018). Gauge-adjusted global satellite mapping of precipitation. *IEEE Transactions on Geoscience and Remote Sensing*, *57*(4), 1928-1935. doi.org/10.1109/tgrs.2018.2870199
32. Militino, A. F., Ugarte, M. D., & Pérez-Goya, U. (2022). Machine learning procedures for daily interpolation of rainfall in Navarre (Spain). *Trends in Mathematical, Information and Data Sciences: A Tribute to Leandro Pardo*, 399-413. doi.org/10.1007/978-3-031-04137-2_34
33. Muhlbauer, A., McCoy, I. L., & Wood, R. (2014). Climatology of stratocumulus cloud morphologies: microphysical properties and radiative effects. *Atmospheric Chemistry and Physics*, *14*(13), 6695-6716. doi.org/10.5194/acp-14-6695-2014
34. Nguyen, G. V., Le, X. H., Van, L. N., Jung, S., Yeon, M., & Lee, G. (2021). Application of random forest algorithm for merging multiple satellite precipitation products across South Korea. *Remote Sensing*, *13*(20), 4033. doi.org/10.3390/rs13204033
35. Pan, B., Hsu, K., AghaKouchak, A., & Sorooshian, S. (2019). Improving precipitation estimation using convolutional neural network. *Water Resources Research*, *55*(3), 2301-2321. doi.org/10.1029/2018wr024090
36. Peña-Arancibia, J. L., Van Dijk, A. I., Renzullo, L. J., & Mulligan, M. (2013). Evaluation of precipitation estimation accuracy in reanalyses, satellite products, and an ensemble method for regions in Australia and South and East Asia. *Journal of Hydrometeorology*, *14*(4), 1323-1333. doi.org/10.1175/jhm-d-12-0132.1
37. Piani, C., Weedon, G. P., Best, M., Gomes, S. M., Viterbo, P., Hagemann, S., & Haerter, J. O. (2010). Statistical bias correction of global simulated daily precipitation and temperature for the application of hydrological models. *Journal of Hydrology*, *395*(3-4), 199-215. doi.org/10.1016/j.jhydrol.2010.10.024
38. Prakash, S., Mitra, A. K., AghaKouchak, A., & Pai, D. S. (2015). Error characterization of TRMM Multisatellite Precipitation Analysis (TMPA-3B42) products over India for different seasons. *Journal of Hydrology*, *529*, 1302-1312. doi.org/10.1016/j.jhydrol.2015.08.062
39. Rajabi Jaghargh, M., Mousavi Baygi, S. M., Araghi, S. A., & Jabari Noghabi, H. (2023). Evaluation of accuracy of daily rainfall values TRMM, GPM, ERA5, and PERSIANN in Razavi Khorasan Province. *Journal of Rainwater Catchment Systems*, 11 (2), 79-101. http://jircsa.ir/article-1-505-fa.html [in Persian].
40. Rata, M., Douaoui, A., Larid, M., & Douaik, A. (2020). Comparison of geostatistical interpolation methods to map annual rainfall in the Chéliff watershed, Algeria. *Theoretical and Applied Climatology*, *141*, 1009-1024. doi.org/10.1007/s00704-020-03218-z
41. Sandri, M., & Zuccolotto, P. (2008). A bias correction algorithm for the Gini variable importance measure in classification trees. *Journal of Computational and Graphical Statistics*, *17*(3), 611-628. doi.org/10.1198/106186008x344522

[DOR: 20.1001.1.24235970.1403.12.1.8.1]

42. Scheel, M. L. M., Rohrer, M., Huggel, C., Santos Villar, D., Silvestre, E., & Huffman, G. J. (2011). Evaluation of TRMM Multi-satellite Precipitation Analysis (TMPA) performance in the Central Andes region and its dependency on spatial and temporal resolution. *Hydrology and Earth System Sciences*, *15*(8), 2649-2663. doi.org/10.5194/hess-15-2649-2011
43. Shen, Z., & Yong, B. (2021). Downscaling the GPM-based satellite precipitation retrievals using gradient boosting decision tree approach over Mainland China. *Journal of Hydrology*, *602*, 126803. doi.org/10.1016/j.jhydrol.2021.126803
44. Sorooshian, S., Duan, Q., & Gupta, V. K. (1993). Calibration of rainfall-runoff models: Application of global optimization to the Sacramento Soil Moisture Accounting Model. *Water Resources Research*, *29*(4), 1185-1194. doi.org/10.1029/92wr02617
45. Sun, Q., Miao, C., Duan, Q., Ashouri, H., Sorooshian, S., & Hsu, K. L. (2018). A review of global precipitation data sets: Data sources, estimation, and intercomparisons. *Reviews of Geophysics*, *56*(1), 79-107. doi.org/10.1002/2017rg000574
46. Tang, T., Chen, T., & Gui, G. (2022). A comparative evaluation of gauge-satellite-based merging products over multiregional complex terrain basin. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *15*, 5275-5287. doi.org/10.1109/jstars.2022.3187983
47. Tao, Y., Gao, X., Hsu, K., Sorooshian, S., & Ihler, A. (2016). A deep neural network modeling framework to reduce bias in satellite precipitation products. *Journal of Hydrometeorology*, *17*(3), 931-945.doi.org/10.1175/jhm-d-15-0075.1
48. Toms, B. A., Barnes, E. A., & Ebert-Uphoff, I. (2020). Physically interpretable neural networks for the geosciences: Applications to earth system variability. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2019MS002002. doi.org/10.1029/2019ms002002
49. Villarini, G., & Krajewski, W. F. (2007). Evaluation of the research version TMPA three-hourly 0.25× 0.25 rainfall estimates over Oklahoma. *Geophysical Research Letters*, *34*(5). doi.org/10.1029/2006gl029147
50. Vrac, M., Noël, T., & Vautard, R. (2016). Bias correction of precipitation through Singularity Stochastic Removal: Because occurrences matter. *Journal of Geophysical Research: Atmospheres*, *121*(10), 5237-5258. doi.org/10.1002/2015jd024511
51. Yang, W., Andréasson, J., Phil Graham, L., Olsson, J., Rosberg, J., & Wetterhall, F. (2010). Distribution-based scaling to improve usability of regional climate model projections for hydrological climate change impacts studies. *Hydrology Research*, *41*(3-4), 211-229. doi.org/10.2166/nh.2010.004
52. Yang, X., Yang, S., Tan, M. L., Pan, H., Zhang, H., Wang, G., ... & Wang, Z. (2022). Correcting the bias of daily satellite precipitation estimates in tropical regions using deep neural network. *Journal of Hydrology*, 608, 127656. doi.org/10.1016/j.jhydrol.2022.127656
53. Yang, Z., Hsu, K., Sorooshian, S., Xu, X., Braithwaite, D., & Verbist, K. M. (2016). Bias adjustment of satellite-based precipitation estimation using gauge observations: A case study in Chile. *Journal of Geophysical Research: Atmospheres*, *121*(8), 3790-3806. doi.org/10.1002/2015jd024540
54. Yuan, F., Wang, B., Shi, C., Cui, W., Zhao, C., Liu, Y., ... & Yang, X. (2018). Evaluation of hydrological utility of IMERG Final run V05 and TMPA 3B42V7 satellite precipitation products in the Yellow River source region, China. *Journal of Hydrology*, *567*, 696-711. doi.org/10.1016/j.jhydrol.2018.06.045
55. Zandi, O., Zahraie, B., Nasseri, M., & Behrangi, A. (2022). Stacking machine learning models versus a locally weighted linear model to generate high-resolution monthly precipitation over a topographically complex area. *Atmospheric Research*, *272*, 106159. doi.org/10.1016/j.atmosres.2022.106159
56. Zhang, J., Fan, H., He, D., & Chen, J. (2019). Integrating precipitation zoning with random forest regression for the spatial downscaling of satellite-based precipitation: A case study of the Lancang–Mekong River Basin. *International Journal of Climatology*, *39*(10), 3947-3961. doi.org/10.1002/joc.6050
57. Zhang, L., Li, X., Zheng, D., Zhang, K., Ma, Q., Zhao, Y., & Ge, Y. (2021). Merging multiple satellite-based precipitation products and gauge observations using a novel double machine learning approach. *Journal of Hydrology*, *594*, 125969. doi.org/10.1016/j.jhydrol.2021.125969
58. Zhang, Y., Zheng, H., Herron, N., Liu, X., Wang, Z., Chiew, F. H., & Parajka, J. (2019). A framework estimating cumulative impact of damming on downstream water availability. *Journal of Hydrology*, *575*, 612-627. doi.org/10.1016/j.jhydrol.2019.05.061
59. Zubieta, R., Getirana, A., Espinoza, J. C., & Lavado, W. (2015). Impacts of satellite-based precipitation datasets on rainfall–runoff modeling of the Western Amazon basin of Peru and Ecuador. *Journal of Hydrology*, *528*, 599-612. doi.org/10.1016/j.jhydrol.2015.06.064

[DOR: 20.1001.1.24235970.1403.12.1.8.1]