Iranian Journal of Soil and Water Research
Homepage: http://ijswr.ut.ac.ir
ISNN: 2423-7833

Reconstruction of Missing Climatic Data Using Combination of Multiple Imputation by Chained Equations (MICE) and Boosting-Based Machine Learning Approaches in the Urmia Lake Basin

Mohammad Shayannejad[1]✉ | **Mohammad Jamali**[2] | **Saeid Eslamian**[3]

1. Corresponding Author, Department of Water Science and Engineering, College of Agriculture, Isfahan University of Technology, Isfahan, Iran. Email: shayannejad@iut.ac.ir
2. Department of Water Science and Engineering, College of Agriculture, Isfahan University of Technology, Isfahan, Iran. Email: mohammad.jamali@ag.iut.ac.ir
3. Department of Water Science and Engineering, College of Agriculture, Isfahan University of Technology, Isfahan, Iran. Email: saeid@iut.ac.ir

Article Info

Article type: Research Article

Article history:

Received: Nov. 18, 2025

Revised: Jan. 21, 2026

Accepted: Apr. 6, 2026

Published online: April. 2026

Keywords:
Boosting learning,
Climatic variables,
MICE algorithm,
Missing data reconstruction,
Urmia Lake Basin

The availability of complete and accurate climatic data plays a crucial role in climatological analyses, hydrological studies, and water resource management. However, meteorological station records often contain missing values, which, if not properly reconstructed, can introduce significant bias into subsequent modeling and analysis. In this study, missing climatic data were reconstructed for six selected meteorological stations—Tabriz, Bonab, Urmia, Maragheh, Saqez, and Sarab—located in the Urmia Lake Basin. Four models, including MICE, MICE–GBR, MICE–XGB, and MICE–LGBM, were developed and compared. Model performance was evaluated using statistical indices such as R^2, NRMSE, |PBIAS|, and KGE. Results revealed that hybrid MICE models based on boosting algorithms provided more accurate and stable reconstructions than the conventional MICE model. Among the tested models, MICE–XGB achieved the best overall performance, with average R^2 exceeding 0.90 and KGE above 0.92 across most stations. The lowest errors were observed for temperature-related variables, while the highest occurred in cloudiness-related parameters. The |PBIAS| values for all models were below 0.025%, indicating negligible systematic bias. Furthermore, model runtime comparisons demonstrated that boosting-based methods, despite their high accuracy, remained computationally efficient and cost-effective. Overall, the findings confirm the superior capability of hybrid MICE models combined with boosting algorithms for reconstructing missing climatic data, highlighting their potential for future climatological and hydrological analyses in data-scarce environments.

Cite this article: Shayannejad, M., Jamali, M, Eslamian, S., (2026) Reconstruction of Missing Climatic Data Using Combination of Multiple Imputation by Chained Equations (MICE) and Boosting-Based Machine Learning Approaches in the Urmia Lake Basin, *Iranian Journal of Soil and Water Research*, 57 (2),467-489. https://doi.org/10.22059/ijswr.2026.406383.670053

Publisher: University of Tehran Press.
DOI: https://doi.org/10.22059/ijswr.2026.406383.670053

EXTENDED ABSTRACT

Introduction

Accurate and continuous climate data are essential for climatological analysis, hydrological modeling, and water resources management, particularly in regions facing rapid climatic fluctuations. However, meteorological records often contain missing values due to sensor malfunctions, network interruptions, and human errors. If these missing data are not properly reconstructed, subsequent analyses such as drought assessment, evapotranspiration modeling, and climate change projections may be biased and unreliable (Afrifa-Yamoah et al., 2020; Hersbach et al., 2020).

Traditional gap-filling techniques such as linear regression, spatial interpolation, and principal component analysis have been widely used, yet they often fail to accurately capture nonlinear and interdependent relationships among climatic variables (Matinzadeh et al., 2013). In recent years, machine learning methods—especially boosting-based algorithms such as Gradient Boosting (GBR), Extreme Gradient Boosting (XGB), and Light Gradient Boosting Machine (LGBM)—have shown superior performance in modeling complex, nonlinear datasets (Alejo-Sanchez et al., 2025).

The present study aims to evaluate and compare the performance of four gap-filling models—MICE, MICE–GBR, MICE–XGB, and MICE–LGBM—for reconstructing missing climate data across six meteorological stations in the Urmia Lake Basin, northwestern Iran. By integrating the multivariate iterative chained equations (MICE) approach with ensemble boosting algorithms, this research investigates whether hybrid learning frameworks can significantly improve reconstruction accuracy and reduce systematic bias in climatic datasets.

Method

The study area includes six representative synoptic stations—Tabriz, Bonab, Urmia, Maragheh, Saqez, and Sarab—with elevations ranging from 1315 to 1682 m. Daily time series of key climatic variables, including temperature (Tmean, Tmin, Tmax), relative humidity (RH), sea-level pressure (SLP), vapor pressure (SVP), cloudiness (CLD), and evapotranspiration (ET), were analyzed. Four gap-filling approaches were developed and implemented in Python using the scikit-learn and XGBoost/LightGBM libraries:

MICE (baseline model): a chained regression-based multiple imputation method;

MICE–GBR: MICE integrated with Gradient Boosting Regression;

MICE–XGB: MICE coupled with Extreme Gradient Boosting;

MICE–LGBM: MICE combined with LightGBM.

The models were evaluated using four widely adopted statistical indices:

Coefficient of Determination (R^2) for accuracy and explained variance;

Normalized Root Mean Square Error (NRMSE) for relative reconstruction error;

Percent Bias (|PBIAS|) for assessing systematic deviation;

Kling–Gupta Efficiency (KGE) for combined evaluation of bias, correlation, and variability.

In addition, the computational runtime of each model was recorded to assess efficiency and potential trade-offs between accuracy and speed.

Results

The comparative analysis demonstrated that the hybrid boosting-based models substantially outperformed the basic MICE model in reconstructing missing climate data. The MICE–XGB model achieved the highest accuracy across most stations and variables, with mean values of $R^2 > 0.90$ and KGE > 0.92. The MICE–LGBM model followed closely, offering comparable performance with slightly lower computational cost. The lowest reconstruction errors were obtained for temperature and pressure-related variables, which exhibit smoother temporal patterns and stronger inter-variable correlations. In contrast, cloudiness (CLD) and evapotranspiration (ET) showed higher error values due to their nonlinear and discontinuous nature. These findings align with Li et al. (2021) and Badrzadeh et al. (2022), who reported similar challenges in reconstructing cloud and radiation data.

Across all models and stations, the absolute percent bias (|PBIAS|) remained below 0.025%, confirming the absence of systematic bias and the robustness of the reconstruction framework. Moreover, spatial evaluation revealed consistent model performance across stations with varying topography and elevation, highlighting the generalizability of the hybrid MICE–boosting methods.

Regarding computational efficiency, the MICE–XGB model, despite being slightly slower than MICE and MICE–LGBM, achieved a favorable balance between accuracy and runtime. Its runtime was less than half that of standard deep learning approaches reported in similar studies, indicating the computational practicality of boosting-based models for large-scale climatic applications.

Conclusions

This study demonstrated that integrating the MICE framework with boosting algorithms such as XGB, GBR, and LGBM significantly enhances the accuracy and reliability of climate data reconstruction. Among the evaluated models, MICE–XGB provided the most consistent and accurate results, particularly for temperature and pressure variables. The extremely low |PBIAS| and high KGE values indicate excellent agreement between reconstructed and observed data, confirming the suitability of these hybrid methods for climatological and hydrological modeling. From a practical perspective, the findings highlight the potential of ensemble-based machine learning approaches in addressing missing data challenges in meteorological datasets—especially in basins such as Urmia Lake, where data discontinuity poses

serious limitations for environmental analysis. The balance between computational efficiency and predictive precision makes these hybrid models ideal candidates for operational climate monitoring systems and regional reanalysis datasets. Future work should focus on extending this framework to spatiotemporal imputation of gridded datasets, integrating remote sensing and reanalysis data, and exploring deep hybrid networks (e.g., MICE–CatBoost or MICE–LSTM) for improved temporal pattern reconstruction.

Funding

The authors received no specific funding for this work.

Authorship contribution

Conceptualization, M.J. and M.Sh.; methodology, M.J. and M.Sh.; software, M.Sh.; validation, M.J., M.Sh. and S.E.; formal analysis, M.J.; investigation, M.J. and M.Sh.; resources, M.J. and S.E.; data curation, M.J.; writing—original draft preparation, M.J.; writing—review and editing, M.Sh. and S.E. All authors have read and agreed to the published version of the manuscript.

Declaration of Generative AI and AI-assisted technologies in the writing process

The authors declare that no generative AI or AI-assisted technologies were used in the writing, analysis, or preparation of this manuscript. The authors take full responsibility for the content of this publication.

Data availability statement

Data available on request from the authors

Acknowledgements

We acknowledge the Iran Meteorological Organization for providing the historical data. The authors thank the anonymous reviewers for their valuable comments and suggestions.

Ethical considerations

The authors avoided data fabrication, falsification, plagiarism, and misconduct.

Conflict of interest

The authors declare no conflict of interest.

مجله تحقیقات آب و خاک ایران، دوره ۵۷، شماره ۲ شاپا: ۷۸۳۳-۲۴۲۳
Homepage: http://ijswr.ut.ac.ir

بازسازی داده‌های گمشده اقلیمی با استفاده از ترکیب روش بازسازی چندگانه با معادلات زنجیره‌ای (MICE) و مدل‌های تقویتی یادگیری ماشین در حوضه آبریز دریاچه ارومیه

محمد شایان‌نژاد[1]✉، محمد جمالی[2]، سعید اسلامیان[3]

۱. نویسنده مسئول، گروه علوم و مهندسی آب، دانشکده کشاورزی، دانشگاه صنعتی اصفهان، اصفهان، ایران. رایانامه: shayannejad@iut.ac.ir

۲. گروه علوم و مهندسی آب، دانشکده کشاورزی، دانشگاه صنعتی اصفهان، اصفهان، ایران. رایانامه: mohammad.jamali@ag.iut.ac.ir

۳. گروه علوم و مهندسی آب، دانشکده کشاورزی، دانشگاه صنعتی اصفهان، اصفهان، ایران. رایانامه: saeid@iut.ac.ir

اطلاعات مقاله

نوع مقاله: مقالهٔ پژوهشی

تاریخ دریافت: ۱۴۰۴/۸/۲۷
تاریخ بازنگری: ۱۴۰۴/۱۱/۱
تاریخ پذیرش: ۱۴۰۵/۱/۱۷
تاریخ انتشار: اردیبهشت ۱۴۰۵

واژه‌های کلیدی:
الگوریتم MICE،
بازسازی داده‌های گمشده،
حوضه آبریز دریاچه ارومیه،
متغیرهای اقلیمی،
یادگیری تقویتی.

چکیده

در دسترس بودن داده‌های کامل و دقیق اقلیمی، نقش کلیدی در تحلیل‌های اقلیم‌شناسی، مطالعات هیدرولوژیکی و مدیریت منابع آب دارد. با این حال، داده‌های ثبت‌شده در ایستگاه‌های هواشناسی معمولاً با گمشدگی مواجه‌اند که در صورت بازسازی نادرست می‌تواند موجب انحراف در نتایج مدل‌سازی شود. در این پژوهش، به‌منظور بازسازی داده‌های گمشده اقلیمی در شش ایستگاه منتخب تبریز، بناب، ارومیه، مراغه، سقز و سراب، واقع در حوضه آبریز دریاچه ارومیه، چهار مدل شامل MICE، MICE–GBR، MICE–XGB و MICE–LGBM مورد بررسی و مقایسه قرار گرفتند. برای ارزیابی کارایی مدل‌ها از شاخص‌های آماری R^2، NRMSE، |PBIAS| و KGE استفاده شد. نتایج نشان داد مدل‌های ترکیبی مبتنی بر الگوریتم‌های تقویتی نسبت به مدل پایه MICE عملکرد دقیق‌تر و باثبات‌تری دارند. در میان آن‌ها، مدل MICE–XGB با میانگین R^2 بالاتر از ۰/۹۰ و KGE بیش از ۰/۹۲ در اغلب ایستگاه‌ها بهترین نتایج را ارائه داد. کمترین خطاها در متغیرهای دمایی و بیشترین در متغیرهای وابسته به ابرناکی مشاهده شد. مقدار |PBIAS| در تمام مدل‌ها کمتر از ۰/۰۲۵ درصد بود که نشان‌دهنده عدم وجود بایاس سیستماتیک قابل توجه است. همچنین مقایسه زمان اجرای مدل‌ها نشان داد روش‌های تقویتی علی‌رغم دقت بالا، از نظر محاسباتی بهینه و مقرون‌به‌صرفه هستند. در مجموع، یافته‌ها بیانگر کارایی بالای مدل‌های ترکیبی MICE با یادگیری تقویتی در بازسازی داده‌های اقلیمی و پیشنهاد به‌کارگیری آن‌ها در تحلیل‌های اقلیمی و هیدرولوژیکی آتی است.

استناد: شایان‌نژاد؛ محمد، جمالی؛ محمد، اسلامیان؛ سعید، (۱۴۰۵|) بازسازی داده‌های گمشده اقلیمی با استفاده از ترکیب روش بازسازی چندگانه با معادلات زنجیره‌ای (MICE)و مدل‌های تقویتی یادگیری ماشین در حوضه آبریز دریاچه ارومیه.، *مجله تحقیقات آب و خاک ایران*، ۵۷ (۲)، ۴۶۷-۴۸۹. https://doi.org/10.22059/ijswr.2026.406383.670053

ناشر: مؤسسه انتشارات دانشگاه تهران.

DOI: https://doi.org/10.22059/ijswr.2026.406383.670053

مقدمه

با توجه به تغییرات روزافزون اقلیم و تشدید نوسانات مکانی و زمانی متغیرهای جوی، دسترسی به داده‌های دقیق، پیوسته و قابل اعتماد از عناصر اقلیمی نظیر دما، بارش، رطوبت نسبی، فشار سطح دریا، تابش خورشیدی و سایر پارامترهای مؤثر، ضرورتی انکارناپذیر در مطالعات علوم زمین به شمار می‌آید (Jamali et al., 2026). این داده‌ها نقش بنیادی در مدل‌سازی اقلیمی، شبیه‌سازی چرخه آب، پیش‌بینی رویدادهای حدی، ارزیابی خشکسالی، مدیریت منابع آب و کشاورزی، و حتی در تصمیم‌گیری‌های زیست‌محیطی و توسعه پایدار دارند (Hersbach et al., 2020; Jamali et al., 2024). با این حال، در عمل، اغلب ایستگاه‌های هواشناسی با مشکل گمشدگی داده‌ها مواجه‌اند. عواملی همچون نقص یا خرابی حسگرها، خطا در ثبت و انتقال اطلاعات، شرایط نامساعد جوی، قطعی‌های ارتباطی، یا حتی تغییر محل و ارتفاع ایستگاه‌ها، سبب ایجاد خلأهای زمانی در سری‌های داده‌ای می‌شوند. در برخی موارد نیز خطاهای انسانی در ثبت و نگهداری داده‌ها باعث حذف یا ثبت نادرست مقادیر می‌گردد. این پدیده در مناطق کوهستانی یا نیمه‌خشک، نظیر شمال‌غرب ایران، به دلیل محدودیت زیرساخت‌های نظارتی، شدیدتر بروز می‌کند. در صورت عدم بازسازی دقیق این داده‌های گمشده، نتایج تحلیل‌ها و مدل‌سازی‌های بعدی مانند تخمین تبخیر و تعرق، مدل‌های بارش–رواناب، یا شاخص‌های خشکسالی ممکن است دچار انحراف و سوگیری شود و به خطاهای معنی‌دار در تصمیم‌گیری‌های مدیریتی منجر گردد (Afrifa-Yamoah et al., 2020). ازاین‌رو، بازسازی داده‌های گمشده اقلیمی نه تنها یک مرحله پیش‌پردازش آماری بلکه بخشی حیاتی از فرآیند تحلیل اقلیم و هیدرولوژی محسوب می‌شود.

هدف اصلی این پژوهش، ارزیابی و مقایسه عملکرد چهار روش بازسازی داده MICE، MICE–GBR، MICE–XGB، و MICE–LGBM برای بازسازی داده‌های اقلیمی گمشده در شش ایستگاه هواشناسی منتخب در حوضه دریاچه ارومیه است. این مطالعه با ترکیب چارچوب چندجایگزینی تکرارشونده MICE که به دلیل حفظ ساختار همبستگی چندمتغیره مورد توجه است با الگوریتم‌های یادگیری تقویتی، به دنبال پر کردن خلاء علمی موجود در زمینه استفاده از رویکردهای ترکیبی قدرتمند برای بهبود دقت بازسازی و کاهش بایاس‌های سیستماتیک در داده‌های اقلیمی است. در حالی که MICE استاندارد از مدل‌های رگرسیونی خطی برای جایگزینی استفاده می‌کند و در روابط غیرخطی محدودیت دارد، استفاده از مدل‌های تقویتی (XGB، GBR، LGBM) امکان مدل‌سازی مؤثر روابط پیچیده و غیرخطی بین متغیرهای اقلیمی را در همان چارچوب MICE فراهم می‌آورد، که این امر انتظار می‌رود پایداری و دقت بازسازی را به میزان قابل توجهی افزایش دهد.

در ادبیات پژوهش، تاکنون روش‌های گوناگونی برای پرکردن داده‌های گمشده ارائه شده است. روش‌های کلاسیک شامل میانگین‌گیری ساده، میان‌یابی خطی و چندجمله‌ای، روش نسبت نرمال، رگرسیون خطی، و تحلیل مؤلفه‌های اصلی (PCA) از متداول‌ترین‌ها هستند (Matinzadeh et al., 2013). اگرچه این روش‌ها در شرایط با درصد گمشدگی کم عملکرد قابل قبولی دارند، اما در مواجهه با روابط غیرخطی میان متغیرها، تعاملات پیچیده اقلیمی یا گمشدگی‌های گسترده، کارایی آن‌ها به‌طور محسوسی کاهش می‌یابد (Alejo-Sanchez et al., 2025). در سال‌های اخیر، با گسترش روش‌های داده‌کاوی و یادگیری ماشین (Machine Learning)، رویکردهای نوینی برای بازسازی داده‌ها مطرح شده‌اند که قادرند الگوهای پیچیده، غیرخطی و چندبعدی موجود در داده‌های اقلیمی را شناسایی کنند. از میان این روش‌ها می‌توان به الگوریتم‌های جنگل تصادفی (Random Forest)، شبکه‌های عصبی مصنوعی (ANN)، ماشین بردار پشتیبان (SVM) و مدل‌های تقویتی (Boosting) اشاره کرد. این مدل‌ها در مقایسه با روش‌های سنتی، توانایی بیشتری در یادگیری ساختارهای درونی داده‌ها دارند و معمولاً دقت و پایداری بالاتری در بازسازی متغیرهای اقلیمی ارائه می‌دهند (Fazel Najafabadi & Shayannejad, 2025). این روش‌ها قادرند الگوهای پیچیده در داده‌ها را بازشناسی کرده و دقت بازسازی را بهبود دهند. در این میان، رویکرد تکمیل چندگانه بر پایه زنجیره معادلات (MICE) به عنوان یکی از روش‌های مؤثر در پرکردن داده‌های گمشده مطرح شده است. این روش با مدل‌سازی تکراری و به‌روزسازی شرطی مقادیر گمشده، امکان بازسازی سازگارتر با ساختار آماری داده‌ها را فراهم می‌کند. با این حال، ترکیب MICE با مدل‌های تقویتی مانند XGBoost، LightGBM و Gradient Boosting در سال‌های اخیر توجه ویژه‌ای را جلب کرده است، زیرا این ترکیب می‌تواند دقت پیش‌بینی را با بهره‌گیری از مزایای یادگیری غیرخطی و بهینه‌سازی گرادیانی بهبود بخشد.

سهم علمی و نوآوری پژوهش حاضر در این مطالعه در سه محور مشخص است: اول، از نظر روش‌شناختی، ترکیب چارچوب چندتعبیره‌ای MICE با الگوریتم‌های تقویتی یادگیری ماشین امکان مدل‌سازی روابط غیرخطی و پیچیده بین متغیرهای اقلیمی را فراهم می‌کند که در مطالعات پیشین به‌طور جامع بررسی نشده بود. دوم، این پژوهش علاوه بر بازسازی داده‌ها در سطح کلی، عملکرد مدل‌ها را در شش ایستگاه منتخب با شرایط اقلیمی متفاوت بررسی می‌کند و دقت، پایداری مکانی و اثر سطح گمشدگی داده‌ها (در این مطالعه

۲۰ درصد) را به‌صورت کمی و بصری تحلیل می‌کند، که خلأ موجود در ارزیابی جامع مدل‌ها بر اساس موقعیت مکانی و نوع متغیر را پر می‌کند. سوم، این مطالعه با ارائه شاخص‌های آماری دقیق و تحلیل حساسیت در سطح ایستگاه‌ها، علاوه بر افزایش دقت بازسازی، امکان برآورد اعتماد و پایداری داده‌های بازسازی‌شده را برای کاربردهای اقلیمی و هیدرولوژیکی فراهم می‌آورد. به این ترتیب، پژوهش حاضر نه تنها خلأ موجود در استفاده از روش‌های ترکیبی قدرتمند برای بازسازی داده‌های اقلیمی را پر می‌کند، بلکه ابزار کمی و بصری برای تحلیل عملکرد مدل‌ها و بررسی ثبات آن‌ها در شرایط مکانی و متغیری متفاوت ارائه می‌دهد و می‌تواند به‌عنوان مرجعی برای بازسازی داده‌های گمشده در حوضه‌های اقلیمی متنوع مورد استفاده قرار گیرد.

پیشینهٔ پژوهش

روش MICE (Multivariate Imputation by Chained Equations) یکی از چارچوب‌های استاندارد و انعطاف‌پذیر برای جایگذاری چندگانه است که با مدل‌سازی شرطی هر متغیر بر اساس سایر متغیرها، مجموعه‌ای از داده‌های کاملِ بازسازی‌شده تولید کرده و عدم قطعیت ناشی از جایگذاری را نیز منعکس می‌سازد (Azur et al., 2011; Van Buuren & Groothuis-Oudshoorn, 2011). مطالعات متعددی نشان داده‌اند که MICE در بسیاری از سناریوها (به‌ویژه هنگامی که درصد گمشدگی کم تا متوسط است و فرض MAR برقرار است) راه‌حل مطمئنی ارائه می‌دهد. Azur et al (2011) کیفیت عملکرد روش MICE را در برآورد داده‌های گمشده در مطالعات روان‌پزشکی بررسی کردند. نتایج این پژوهش نشان داد که روش MICE در شرایطی که داده‌های مفقود تحت فرض MAR باشند و نرخ مفقودیت در محدوده پایین تا متوسط قرار گیرد، تخمین‌هایی پایدار، باثبات و نزدیک به مقادیر واقعی ارائه می‌دهد. Golkhatmi and Farzandi (2024) با هدف کنترل کیفی و بازسازی داده‌های ناقص بارش در حوضه قره‌قوم شمال‌شرقی ایران، از داده‌های ۱۴۱ ایستگاه استفاده کردند. در این پژوهش ابتدا خطاهای درشت، ناسازگاری زمانی و داده‌های پرت مورد بررسی قرار گرفت. سپس اطلاعاتی نظیر تعداد روزهای بارش در هر ماه، بیشینه بارش ماهانه و بارش استانداردشده برای شناسایی آلودگی داده‌های پرت به‌کار گرفته شد. پژوهشگران برای بازسازی داده‌های ناقص، از بسته‌ی MICE در نرم‌افزار R استفاده کردند که داده‌های مفقود را با استفاده از زنجیره‌ای از معادلات برآورد می‌کند. پنج تابع مختلف این بسته مورد ارزیابی قرار گرفت و نتایج نشان داد که روش norm.nob دقیق‌ترین برآورد را ارائه داده و روش‌های sample و mean عملکرد ضعیف‌تری داشتند. یافته‌های این تحقیق نشان داد که ترکیب کنترل کیفی داده‌ها با روش MICE می‌تواند رویکردی کارآمد برای بازسازی داده‌های ناقص در پژوهش‌های اقلیمی و هیدرولوژیکی باشد، به‌ویژه در شرایطی که نرخ مفقودیت پایین تا متوسط بوده و فرض MAR برقرار است (Costa et al., 2024).

Plein et al (2025) در پژوهشی با هدف بازسازی داده‌های ناقص دما و رطوبت در شبکه ایستگاه‌های هواشناسی شهر فرایبورگ آلمان، از روش Extreme Gradient Boosting (XGBoost) استفاده کردند. داده‌های دارای شکاف‌های مصنوعی ۱ تا ۲۸ روزه مورد ارزیابی قرار گرفت و نتایج نشان داد که مدل XGBoost در بازسازی داده‌ها دقت بالایی دارد، به‌طوری‌که میانگین خطای RMSE برای دما ۰/۴۶ کلوین و برای رطوبت نسبی ۲/۵۱ درصد بود. عملکرد مدل نسبت به طول شکاف حساسیت کمی داشت اما در ایستگاه‌های دور از مناطق شهری دقت کمتر بود. داده‌های بازسازی‌شده نشان دادند که مرکز شهر به‌طور میانگین ۱/۱ درجه گرم‌تر از مناطق روستایی بوده و فشار بخار در آن ۷ درصد کمتر است. این پژوهش بیانگر کارایی بالای مدل‌های یادگیری تقویتی مانند XGBoost در بازسازی داده‌های ناقص شبکه‌های هواشناسی شهری است. Jääskeläinen et al (2022) به‌منظور بازسازی داده‌های ناقص آلبیدوی سطحی یخ‌های دریایی قطب شمال از روش گرادیان بوستینگ استفاده کردند. داده‌های ۳۴ ساله به‌دلیل پوشش ابری و زاویه زیاد تابش، دارای مقادیر مفقود بودند. مدل گرادیان بوستینگ با بهره‌گیری از میانگین ماهانه آلبیدو، دمای روشنایی و غلظت یخ دریا آموزش داده شد. نتایج نشان داد این روش با میانگین خطای RMSE حدود ۰/۰۴۸ و اختلاف نسبی کمتر از ۲۰ درصد، عملکردی دقیق‌تر و باثبات‌تر از مدل‌های رگرسیون خطی دارد و در پیش‌بینی آلبیدوی یخ‌های در حال ذوب نیز نتایج بهتری ارائه می‌دهد. در بسیاری از مطالعات اخیر نیز تأکید شده است که ترکیب روش‌های چندتکمیلی و الگوریتم‌های یادگیری ماشین می‌تواند به عنوان راهی نوین برای مقابله با چالش پرکردن داده‌های اقلیمی مطرح شود (Alejo-Sanchez et al., 2025; Costa et al., 2024; Jamali et al., 2026; Hosseinpour et al., 2025).

هدف اصلی این پژوهش، ارزیابی و مقایسه عملکرد چهار رویکرد بازسازی داده‌های اقلیمی شامل روش استاندارد MICE و سه مدل ترکیبی MICE–GBR، MICE–XGB و MICE–LGBM در مواجهه با داده‌های گمشده شش ایستگاه منتخب حوضه دریاچه ارومیه (تبریز، بناب، ارومیه، مراغه، سقز و سراب) در شمال‌غرب ایران است. در این مطالعه، چارچوب MICE به‌عنوان بنیان آماری اصلی به‌کار

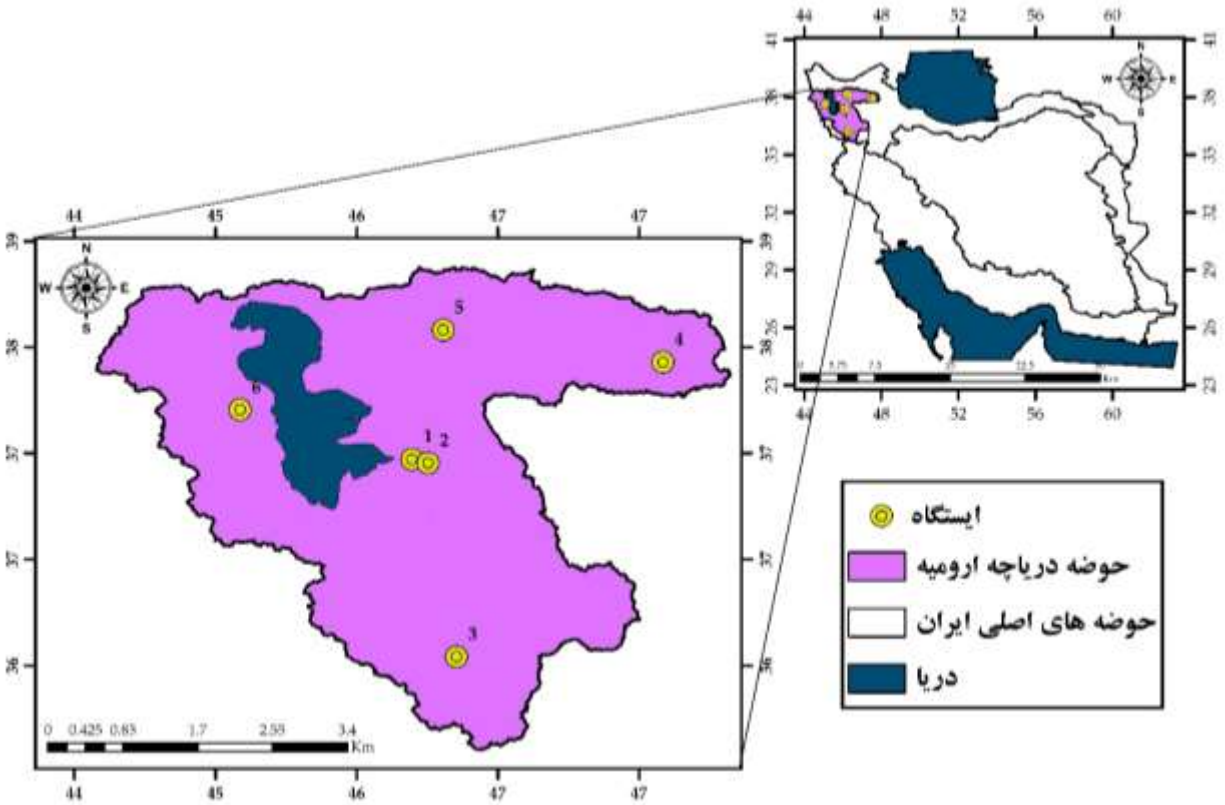
گرفته شده است که با بهره‌گیری از راهبرد چندجایگزینی تکرارشونده، امکان حفظ ساختار همبستگی‌ها و ویژگی‌های آماری چندمتغیره داده‌های اقلیمی را فراهم می‌کند. با این حال، روش‌های کلاسیک مورد استفاده در پیاده‌سازی سنتی MICE، نظیر رگرسیون خطی، در بازنمایی روابط غیرخطی و پیچیده حاکم بر متغیرهای اقلیمی منطقه‌ای با محدودیت‌هایی مواجه‌اند. ازاین‌رو، پژوهش حاضر با ادغام چارچوب MICE و الگوریتم‌های تقویتی پیشرفته که در مطالعات بین‌المللی کارایی بالای خود را در مدل‌سازی غیرخطی نشان داده‌اند، درصدد رفع این کاستی است. هدف نهایی، بررسی این موضوع است که آیا چارچوب‌های ترکیبی پیشنهادی قادرند بر چالش‌های بازسازی داده در شرایط اقلیمی پیچیده و ناهمگن منطقه‌ای غلبه کرده و دقت و پایداری بازسازی را به‌طور معناداری بهبود دهند. در مجموع، این پژوهش نخستین تلاش نظام‌مند برای تلفیق مدل‌های یادگیری ماشین تقویتی با چارچوب MICE در حوضه دریاچه ارومیه به‌شمار می‌رود و می‌کوشد رویکردی دقیق، پایدار و کارآمد برای مدیریت و بازسازی داده‌های اقلیمی در این منطقه حساس ارائه دهد.

با این حال، باید اذعان داشت که فرض تصادفی بودن داده‌های گمشده و محدود بودن مطالعه به یک حوضه خاص، می‌تواند تعمیم‌پذیری نتایج را تا حدی محدود سازد و همچنین حساسیت روش‌های تقویتی به تنظیم دقیق پارامترها، از دیگر چالش‌های این پژوهش محسوب می‌شود.

روش‌شناسی پژوهش

منطقه مورد مطالعه

حوضه دریاچه ارومیه به عنوان بزرگ‌ترین حوضه بسته[1] ایران و یکی از اکوسیستم‌های منحصر به فرد خاورمیانه، در شمال‌غرب کشور واقع شده است. این حوضه نه تنها از نظر زیست‌محیطی و اکولوژیک حائز اهمیت است، بلکه از دیدگاه اقتصادی و اجتماعی نیز نقش کلیدی در معیشت میلیون‌ها نفر از ساکنان استان‌های آذربایجان شرقی، آذربایجان غربی و کردستان ایفا می‌کند. در دهه‌های اخیر، کاهش شدید تراز آب دریاچه و بحران خشکسالی، اهمیت پایش و تحلیل داده‌های اقلیمی این منطقه را بیش از پیش برجسته ساخته است.



شکل ۱. نقشه حوضه آبریز دریاچه ارومیه و محل ایستگاه‌های منتخب.

برای انجام این پژوهش، شش ایستگاه سینوپتیک کلیدی و با خصوصیات فیزیکی و پراکندگی مکانی مختلف شامل تبریز، بناب،

1. Endorheic Basin

ارومیه، مراغه، سقز و سراب انتخاب شدند (شکل ۱ و جدول ۱). این ایستگاه‌ها در نقاط مختلف حوضه پراکنده‌اند و به دلیل تفاوت در موقعیت جغرافیایی، ارتفاع و شرایط اقلیمی، طیفی از شرایط آب‌وهوایی منطقه را نمایندگی می‌کنند. این تنوع مکانی و اقلیمی، بستر مناسبی برای ارزیابی عملکرد روش‌های بازسازی داده در شرایط مختلف فراهم می‌آورد.

جدول ۱. مشخصات ایستگاه‌های منتخب در پژوهش.

شماره	ایستگاه	طول جغرافیایی (°E)	عرض جغرافیایی (°N)	ارتفاع (m)	سال شروع	سال پایان	دوره آماری
۱	بناب	۴۶/۰۵۲	۳۷/۳۷۰	۱۲۹۰	۱۹۹۹	۲۰۲۴	۲۶
۲	مراغه	۴۶/۱۴۶	۳۷/۳۴۸	۱۴۷۷/۷	۱۹۸۶	۲۰۲۴	۳۹
۳	سقز	۴۶/۳۱۱	۳۶/۲۲۱	۱۵۲۲/۸	۱۹۶۱	۲۰۲۴	۶۴
۴	سراب	۴۷/۵۰۸	۳۷/۹۳۵	۱۶۸۲	۱۹۸۶	۲۰۲۴	۳۹
۵	تبریز	۴۶/۲۳۴	۳۸/۱۲۲	۱۳۶۱	۱۹۵۱	۲۰۲۴	۷۴
۶	ارومیه	۴۵/۰۵۶	۳۷/۶۵۸	۱۳۱۵/۹	۱۹۵۱	۲۰۲۴	۷۴

داده‌های مورد استفاده

در این پژوهش، ۱۴ متغیر اقلیمی روزانه به‌عنوان داده‌های پایه انتخاب و بررسی شدند (جدول ۲). این متغیرها شامل پارامترهای بارش، دما، رطوبت، فشار، تابش، تبخیر و تعرق، ابرناکی و باد بوده و به‌طور جامع ابعاد مختلف شرایط اقلیمی حاکم بر حوضه آبریز دریاچه ارومیه را پوشش می‌دهند. بررسی داده‌های خام نشان داد که تمامی متغیرها دارای مقادیر گمشده در بازه‌های زمانی مختلف بوده‌اند. ازاین‌رو، فرآیند بازسازی داده‌ها برای همه متغیرهای اقلیمی به‌صورت مستقل و با بهره‌گیری از روش MICE و همچنین رویکردهای تقویتی یادگیری ماشین (Gradient Boosting (GBR), Extreme Gradient Boosting (XGB)) و Light Gradient Boosting Machine (LGBM)) اجرا گردید. داده‌های اقلیمی مورد استفاده از ایستگاه‌های سینوپتیک واقع در سطح حوضه گردآوری شده (جدول ۱) و به‌صورت روزانه در دسترس قرار گرفتند. قبل از اجرای فرآیند بازسازی، داده‌ها تحت کنترل کیفی قرار گرفته و مقادیر پرت و ناهنجاری‌ها حذف یا اصلاح شدند تا کیفیت داده‌های ورودی برای مدل‌سازی تضمین شود.

در این مطالعه، سازوکار گمشدگی داده‌ها از نوع Missing At Random (MAR) فرض شده است؛ به این معنا که احتمال گمشده بودن مشاهدات می‌تواند به داده‌های مشاهده‌شده وابسته باشد، اما به مقادیر واقعی داده‌های گمشده وابستگی ندارد. این فرض با توجه به ماهیت داده‌های اقلیمی و محدودیت‌های ثبت و گردآوری داده‌ها، به‌عنوان یک تقریب مناسب در نظر گرفته شده و مبنای به‌کارگیری روش MICE قرار گرفته است.

جدول ۲. متغیرهای اقلیمی روزانه منتخب در این پژوهش.

نماد	نام متغیر	واحد	نماد	نام متغیر	واحد
Tmax	بیشینه دما	°C	CLDmean	میانگین ابرناکی	oktas
Tmin	کمینه دما	°C	CLDmax	بیشینه ابرناکی	oktas
Tmean	میانگین دما	°C	RHmax	بیشینه رطوبت هوا	%
TDmean	میانگین دمای نقطه شبنم	°C	RHmin	کمینه رطوبت هوا	%
SLPmean	میانگین فشار سطح دریا	hPa	RHmean	میانگین رطوبت هوا	%
VPmean	میانگین فشار بخار آب	hPa	Sun	ساعت آفتابی	h
SVPmean	میانگین فشار بخار اشباع	hPa	ET	تبخیر و تعرق	mm

پیش‌پردازش داده‌ها

پیش از اجرای فرآیند بازسازی داده‌های گمشده، داده‌های اقلیمی جمع‌آوری‌شده از ایستگاه‌های منتخب تحت مراحل پیش‌پردازش اولیه قرار گرفتند تا کیفیت و یکپارچگی سری‌های زمانی ارزیابی شود. در این مرحله، داده‌ها از نظر وجود مقادیر گمشده، مقادیر پرت و ناسازگاری‌های زمانی بررسی شدند. مقادیر گمشده موجود در سری‌های ثبت‌شده صرفاً شناسایی و مستندسازی شدند تا میزان و الگوی گمشدگی داده‌ها در هر ایستگاه و برای هر متغیر مشخص شود، بدون آن‌که در این مرحله اقدامی برای جایگزینی یا بازسازی آن‌ها صورت گیرد. شناسایی مقادیر پرت با استفاده از حدود آماری و نمودارهای جعبه‌ای انجام شد و تنها در مواردی که خطاهای آشکار ناشی از ثبت

یا انتقال داده‌ها تشخیص داده شد، این مقادیر حذف و به‌عنوان داده‌های گمشده در نظر گرفته شدند تا در مرحله بازسازی، تحت فرض MAR و با استفاده از روش MICE مدیریت شوند.

به‌منظور اطمینان از همسانی داده‌ها، تمامی سری‌های زمانی هر ایستگاه به‌صورت مستقل به فرمت روزانه تبدیل شدند. این مرحله صرفاً شامل یکنواخت‌سازی زمانی بوده و هیچ‌گونه نرمال‌سازی آماری بین ایستگاه‌ها اعمال نشده است. داده‌های ناقص یا با فرمت نامناسب اصلاح و به قالب یکنواخت تبدیل گردیدند. در این مرحله، متغیرها به صورت جداگانه پردازش شدند تا آماده اعمال روش‌های بازسازی گردند. برای ارزیابی عملکرد روش MICE و مدل‌های یادگیری ماشین، ۲۰ درصد از داده‌ها به صورت عمدی و تصادفی حذف شدند تا شرایط گمشدگی شبیه‌سازی گردد. انتخاب سطح ۲۰ درصد مبتنی بر دو دلیل عمده است: نخست، مطالعات نشان داده‌اند که در صورتی که گمشدگی کمتر از ۱۰ درصد باشد، بسیاری از روش‌های ساده نیز عملکردی قابل‌قبول دارند؛ اما با افزایش به بازه ۱۰ تا ۳۰ درصد، تمایز میان روش‌های پیشرفته و ساده آشکارتر می‌شود و روش‌هایی مانند برتری خود را نشان می‌دهند (Azur et al., 2011; Schafer & Graham, 2002). دوم، بررسی داده‌های اقلیمی ایران نشان می‌دهد که میزان داده‌های ناقص در ایستگاه‌ها معمولاً بین ۱۵ تا ۲۵ درصد متغیر است (Farzandi et al., 2022; Hasanpour Kashani & Dinpashoh, 2012; Khosravi et al., 2015)، بنابراین انتخاب سطح ۲۰ درصد همسو با شرایط واقعی شبکه‌های اقلیم‌سنجی کشور و به‌ویژه حوضه دریاچه ارومیه است.

پس از ایجاد گمشدگی مصنوعی، داده‌ها به دو مجموعه تقسیم شدند:

مجموعه بازسازی: که شامل ۲۰ درصد داده‌های ناقص برای اعمال روش MICE و مدل‌های یادگیری ماشین.

مجموعه ارزیابی: شامل داده‌های واقعی حذف شده که به عنوان مرجع برای سنجش دقت بازسازی استفاده شدند.

علاوه بر ارزیابی دقت بازسازی، عدم‌قطعیت نتایج نیز مورد بررسی قرار گرفت. بدین منظور، فرآیند بازسازی داده‌ها در چارچوب اعتبارسنجی متقابل انجام شد؛ به‌طوری‌که در هر تکرار، الگوی گمشدگی و داده‌های آموزشی–آزمون متفاوتی مورد استفاده قرار گرفت. تغییرات مقادیر شاخص‌های عملکرد (از جمله NRMSE، R^2 و KGE) در تکرارهای مختلف به‌عنوان معیاری از پایداری و عدم‌قطعیت بازسازی در نظر گرفته شد. این رویکرد امکان ارزیابی حساسیت نتایج بازسازی نسبت به نحوه انتخاب داده‌های مرجع و الگوی گمشدگی را فراهم می‌کند.

متغیرها به شکل ماتریسی سازماندهی شدند که شامل مقادیر روزانه هر متغیر در هر ایستگاه بود. این قالب، امکان اعمال الگوریتم MICE و مدل‌های Gradient Boosting Regressor، XGBoost و LightGBM را به صورت یکپارچه فراهم کرد و همچنین اجازه داد تا عملکرد بازسازی در متغیرهای مختلف و ایستگاه‌های متفاوت مقایسه گردد.

روش‌های بازسازی داده

در این پژوهش، بازسازی داده‌های گمشده اقلیمی با بهره‌گیری از دو رویکرد اصلی (مطابق جدول ۳) انجام شد:

۱- روش آماری بازسازی چندگانه با معادلات زنجیره‌ای (MICE)

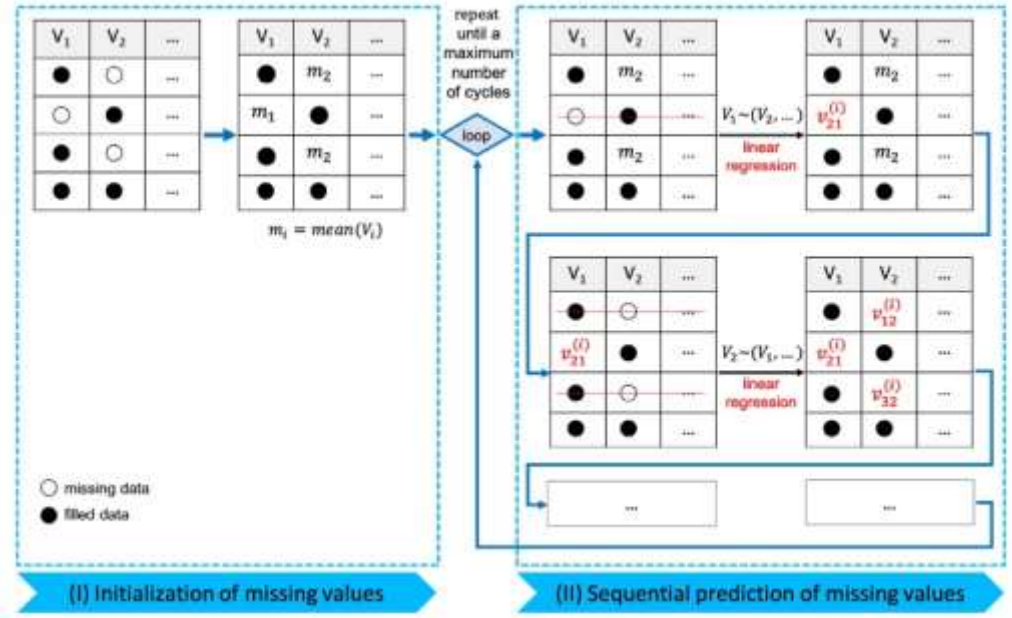
۲- ترکیب با الگوریتم‌های یادگیری ماشین تقویتی (Gradient Boosting (GBR), Extreme Gradient Boosting (XGB) و Light Gradient Boosting Machine (LGBM))

جدول ۳. مدل‌های مورد استفاده برای بازسازی داده‌های گمشده در این پژوهش.

دسته بندی	نماد	نام مدل	توضیحات
آماری	MICE	Multiple Imputation by Chained Equations	روش بازسازی چندگانه با معادلات زنجیره‌ای، بازسازی داده‌های چندمتغیره با حفظ ساختار همبستگی
ترکیبی	MICE-GBR	MICE + Gradient Boosting Regressor	ترکیب MICE با الگوریتم گرادیان بوستینگ رگرسیونی، مدل‌سازی روابط غیرخطی و بهبود بازسازی داده‌های عددی.
	MICE-XGB	MICE + Extreme Gradient Boosting	ترکیب MICE با نسخه بهینه و سریع گرادیان بوستینگ با منظم‌سازی و پردازش موازی.
	MICE-LGBM	MICE + Light Gradient Boosting Machine	ترکیب MICE با الگوریتم بوستینگ کارآمد Leaf-wise مناسب برای داده‌های بزرگ و پرابعاد.

۱- روش MICE

روش یکی از پرکاربردترین و معتبرترین روش‌ها برای بازسازی داده‌های گمشده در مطالعات آماری و علوم زیست‌محیطی است (Little & Rubin, 1987; Van Buuren, 2000). این روش بر مبنای رویکرد «چندجایگزینی» عمل می‌کند؛ به این معنا که برای هر داده گمشده چند مقدار جایگزین بر اساس توزیع شرطی سایر متغیرها تولید می‌شود. فرآیند در قالب یک زنجیره از معادلات رگرسیونی تکرارشونده اجرا می‌شود و در هر گام، داده‌های گمشده یک متغیر با توجه به سایر متغیرها تخمین زده می‌شود (شکل ۲).



شکل ۲. فلوچارت روش MICE

از مهم‌ترین مزایای می‌توان به موارد زیر اشاره کرد:

قابلیت استفاده در داده‌های چندمتغیره و پیچیده،

حفظ همبستگی‌ها و ساختار وابستگی بین متغیرها،

انعطاف‌پذیری در انتخاب مدل رگرسیونی مناسب برای هر متغیر (خطی، لجستیک و ...)،

امکان برآورد عدم‌قطعیت ناشی از جایگزینی داده‌ها از طریق تولید چند مجموعه داده تکمیل‌شده.

با وجود این مزایا، کارایی در شرایطی که روابط بین متغیرها غیرخطی یا بسیار پیچیده باشند، ممکن است محدود شود و دقت بازسازی کاهش یابد (Azur et al., 2011). از این‌رو، ترکیب آن با الگوریتم‌های یادگیری ماشین پیشرفته می‌تواند بهبود چشمگیری در نتایج ایجاد کند.

۲- ترکیب روش MICE با الگوریتم‌های تقویتی

در سال‌های اخیر، الگوریتم‌های یادگیری ماشین تقویتی به‌عنوان ابزارهای قدرتمندی برای مدل‌سازی روابط پیچیده و غیرخطی در داده‌های اقلیمی و محیطی معرفی شده‌اند (Chen & Guestrin, 2016; Davari et al., 2025; Ke et al., 2017). در این پژوهش، سه الگوریتم تقویتی در کنار مورد استفاده قرار گرفتند:

این الگوریتم بر اساس یادگیری تدریجی درخت‌های تصمیم ضعیف عمل می‌کند. هر مدل جدید خطاهای مدل قبلی را اصلاح کرده و در نهایت یک مدل قوی و پایدار ایجاد می‌شود (Friedman, 2001).

نسخه بهینه‌شده گرادیان بوستینگ است که با منظم‌سازی و ، جلوگیری از بیش‌برازش و قابلیت پردازش موازی، سرعت و دقت بالاتری دارد (Chen & Guestrin, 2016).

الگوریتم LightGBM که الگوریتمی سبک و سریع برای یادگیری ماشین تقویتی محسوب می‌شود. این الگوریتم با رشد درخت‌ها به‌صورت Leaf-wise و استفاده بهینه از حافظه، امکان پردازش مجموعه داده‌های بزرگ و با ابعاد بالا را فراهم می‌آورد و زمان آموزش نسبت به الگوریتم‌های سنتی Gradient Boosting و XGBoost کاهش می‌یابد. این ویژگی‌ها باعث می‌شود LightGBM مناسب

کاربردهای محاسباتی سنگین و داده‌های بزرگ اقلیمی باشد (Ke et al., 2017).

ترکیب این الگوریتم‌ها با به‌گونه‌ای طراحی شد که فرآیند چندجایگزینی داده‌ها (به‌عنوان بخش آماری) با توانایی الگوریتم‌های یادگیری ماشین در مدل‌سازی غیرخطی ادغام گردد. این ترکیب مزیت دوگانه‌ای ایجاد می‌کند:

۱. حفظ ویژگی‌های آماری و ساختار همبستگی داده‌ها توسط،

۲. افزایش دقت بازسازی در شرایط داده‌های اقلیمی پیچیده توسط الگوریتم‌های تقویتی.

در این پژوهش، الگوریتم‌های تقویتی به‌عنوان مدل‌های جایگزینی (Imputation Models) در چارچوب MICE به‌کار گرفته شدند. به‌عبارت دیگر، برای هر متغیر گمشده، یک مدل تقویتی آموزش داده می‌شود تا مقادیر گمشده آن متغیر بر اساس سایر متغیرهای مشاهده‌شده پیش‌بینی شود. فرآیند چندجایگزینی (Multiple Imputation) در چارچوب MICE انجام شده و چندین مجموعه داده بازسازی‌شده مستقل تولید شده‌اند. سپس مقادیر پیش‌بینی‌شده از همه مجموعه‌ها تجمیع شده و شاخص‌های عملکرد مدل‌ها (NRMSE، R^2 و KGE) محاسبه گردید تا هم دقت بازسازی و هم عدم‌قطعیت فرآیند بازسازی منعکس شود. این رویکرد، تلفیق قدرت آماری MICE در حفظ ساختار همبستگی با توانایی الگوریتم‌های یادگیری ماشین در مدل‌سازی روابط غیرخطی را به‌طور همزمان ممکن می‌سازد.

مطالعات اخیر نیز کارایی چنین رویکردهای ترکیبی را در حوزه اقلیم و هیدرولوژی نشان داده‌اند؛ برای مثال، Farzandi et al (2022) در ایران نشان داده‌اند که روش‌های ترکیبی می‌توانند عملکردی به‌مراتب بهتر از رویکردهای کلاسیک ارائه دهند.

رویکرد اعتبارسنجی[1]

برای ارزیابی عملکرد مدل‌ها، از روش اعتبارسنجی متقابل پنج‌تایی استفاده شد. در این رویکرد، داده‌های آموزشی به پنج زیرمجموعه مساوی تقسیم گردید و در هر تکرار، چهار زیرمجموعه برای آموزش مدل و یک زیرمجموعه برای اعتبارسنجی به کار گرفته شد. این فرآیند پنج بار تکرار شد به‌طوری‌که هر زیرمجموعه یک بار نقش داده‌های اعتبارسنجی را ایفا کند. در نهایت، میانگین نتایج حاصل از پنج مرحله به عنوان معیار نهایی عملکرد مدل گزارش شد که برآوردی پایدار و قابل اعتماد از توانایی مدل فراهم ساخت. به منظور سنجش دقت بازسازی داده‌های گمشده نیز مقادیر واقعی حذف‌شده به‌طور عمدی به عنوان مرجع مورد استفاده قرار گرفتند و مقادیر بازسازی‌شده توسط مدل‌ها با این مقادیر واقعی مقایسه شدند. این مقایسه امکان بررسی هم‌زمان اثر حذف مصنوعی داده‌ها بر کیفیت بازسازی و توانایی مدل‌ها در برآورد مقادیر واقعی تحت شرایط وجود داده‌های ناقص را فراهم کرد. رویکرد انتخاب‌شده چندین مزیت کلیدی داشت؛ از جمله کاهش احتمال بیش‌برازش مدل‌ها بر روی داده‌های آموزش، ایجاد شرایطی نزدیک به وضعیت واقعی که در آن داده‌های اقلیمی اغلب ناقص هستند، و فراهم آوردن امکان مقایسه مستقیم میان روش‌ها تحت شرایط یکسان.

انتخاب اعتبارسنجی پنج‌تایی، در مقایسه با روش‌های ساده‌تر مانند تفکیک یک‌باره داده‌ها و نیز روش‌های پرهزینه‌تری مانند یا اعتبارسنجی با تعداد تاخوردگی بالاتر، توازنی مناسب میان دقت برآورد خطا، پایداری نتایج و هزینه محاسباتی ایجاد می‌کند. این رویکرد به‌ویژه برای داده‌های اقلیمی با حجم محدود و ساختار همبسته زمانی، روشی متداول و قابل اعتماد برای ارزیابی عملکرد و کاهش بیش‌برازش مدل‌ها محسوب می‌شود.

شاخص‌های ارزیابی

به‌منظور سنجش دقت و کارایی روش‌های بازسازی داده‌های گمشده، از چند شاخص آماری معتبر و پرکاربرد استفاده شد که ابعاد مختلف عملکرد مدل‌ها از جمله دقت، بایاس و تغییرپذیری را ارزیابی می‌کنند. شاخص‌های مورد استفاده شامل ضریب تعیین (R^2)، خطای جذر میانگین مربعات نرمال‌شده (NRMSE)، درصد بایاس (PBIAS) و کارایی کلینگ–گوپتا (KGE) هستند. در ادامه، فرمول و تفسیر هر یک از شاخص‌ها ارائه می‌شود.

۱- ضریب تعیین[2] (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$ رابطه ۱)

که در آن y_i و $\hat{y}_i$ به ترتیب نشان‌دهنده مقادیر مشاهده‌شده و پیش‌بینی‌شده و تعداد کل مشاهدات است. این شاخص میزان همبستگی خطی بین مقادیر بازسازی‌شده و مقادیر واقعی را نشان می‌دهد. مقدار آن در بازه ۰ تا ۱ قرار دارد؛ مقادیر نزدیک به ۱ بیانگر برازش

1. Cross-Validation Approach
2. Coefficient of Determination

مطلوب‌تر مدل است (Legates & McCabe Jr, 1999; Willmott, 1981).

۲- خطای جذر میانگین مربعات نرمال‌شده[1](NRMSE)

رابطه ۲) $$NRMSE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}}{\bar{y}}$$

شاخص هنگام تقسیم بر میانگین مقادیر واقعی، همواره مقداری غیرمنفی دارد و هرچه به صفر نزدیک‌تر باشد، دقت پیش‌بینی مدل بالاتر است، و مقادیر بزرگ‌تر نشان‌دهنده خطای نسبی بیشتر نسبت به میانگین داده‌ها هستند (Jamali Jezeh et al., 2020; Jamali et al., 2026; Legates & McCabe Jr, 1999).

۳- درصد بایاس (PBIAS)

رابطه ۳) $$PBIAS = \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)}{\sum_{i=1}^{n}y_i} \times 100$$

شاخص درصد بایاس (PBIAS) میزان تمایل سیستماتیک خطای مدل را نشان می‌دهد؛ مقدار صفر بیانگر پیش‌بینی بدون بایاس است، مقادیر مثبت نشان‌دهنده پیش‌بینی بیش از حد و مقادیر منفی نشان‌دهنده پیش‌بینی کمتر از مقادیر واقعی می‌باشد، به‌طوری که PBIAS در حدود ±۱۰ درصد کیفیت عالی، ۲۰–±۱۰ درصد کیفیت خوب، ۳۰–±۲۰ درصد کیفیت متوسط و مقادیر بیش از ±۳۰ درصد کیفیت ضعیف مدل را نشان می‌دهد (Gupta Hoshin et al., 1999; N. Moriasi et al., 2007).

۴- کارایی کلینگ–گوپتا[2](KGE)

معیاری جامع برای ارزیابی عملکرد مدل‌های هیدرولوژیکی است که همزمان سه جنبه مهم پیش‌بینی مدل یعنی همبستگی، بایاس و تغییرپذیری را در نظر می‌گیرد (Gupta et al., 2009; Knoben et al., 2019).

رابطه ۴) $$KGE = 1 - \sqrt{(1-r)^2 + (1-\beta)^2 + (1-\gamma)^2}$$

که در آن: r ضریب همبستگی بین مقادیر پیش‌بینی و واقعی، $\beta = \frac{\bar{y}_m}{\bar{y}_o}$ نسبت میانگین پیش‌بینی به میانگین واقعی (نمایانگر بایاس) و $\gamma = \frac{CV_m}{CV_o}$ نسبت ضریب تغییرات پیش‌بینی به واقعی (نمایانگر تغییرپذیری) است.

مقدار ایده‌آل شاخص برابر ۱ است و هرچه مقدار آن به ۱ نزدیک‌تر باشد، عملکرد مدل مطلوب‌تر خواهد بود.

یافته‌های پژوهش

بررسی کلی عملکرد روش‌ها

به‌منظور ارزیابی جامع کارایی روش‌های مختلف بازسازی داده، شاخص‌های آماری R^2، NRMSE، |PBIAS| و KGE برای چهار مدل MICE، MICE–GBR، MICE–XGB و MICE–LGBM در شش ایستگاه منتخب حوضه آبریز دریاچه ارومیه محاسبه شدند (جدول ۴).

در این جدول مقادیر کمینه و بیشینه هر شاخص به همراه متغیر اقلیمی متناظر با آن مقدار نمایش داده شده است تا محدوده تغییرات و رفتار مدل‌ها در بازسازی متغیرهای اقلیمی مختلف به روشنی مشخص شود. بر اساس نتایج، مدل‌های ترکیبی مبتنی بر یادگیری تقویتی به‌طور کلی عملکرد بهتری نسبت به روش پایه MICE نشان دادند، هرچند نتایج بازسازی در هر چهار روش از دقت قابل قبول و مطلوبی برخوردار بوده است. در میان مدل‌های مورد بررسی، مدل MICE–XGB بیشترین مقدار R^2 را در بیشتر ایستگاه‌ها به خود اختصاص داد؛ به‌گونه‌ای که مقدار این شاخص در اغلب متغیرها بالاتر از ۰/۹۰ برآورد شد. پس از آن، مدل MICE–LGBM نیز دقت قابل‌توجهی در بازسازی متغیرهای دمایی و فشاری از خود نشان داد. بیشترین مقدار R^2 (۰/۹۹۶) در ایستگاه ارومیه با مدل MICE–XGB و برای متغیر میانگین دمای روزانه (Tmean) مشاهده شد، در حالی‌که کمترین مقدار آن (۰/۰۲۶) در ایستگاه بناب با مدل MICE و برای متغیر میانگین فشار سطح دریا (SLPmean) به‌دست آمد. به‌طور کلی، متغیرهای دمایی به‌ویژه میانگین دمای روزانه (Tmean) بیشترین مقادیر R^2 را در تمامی ایستگاه‌ها به خود اختصاص داده‌اند، در حالی‌که کمترین مقادیر R^2 مربوط به میانگین فشار سطح دریا (SLPmean) بوده است. دلیل این امر آن است که متغیرهای دمایی معمولاً دارای نوسانات منظم‌تر و همبستگی مکانی–زمانی بالاتری هستند، در حالی‌که فشار سطح دریا تحت تأثیر عوامل سینوپتیکی و دینامیکی بزرگ‌مقیاس قرار دارد و رفتار ناپیوسته‌تری نسبت به سایر متغیرها دارد؛ بنابراین مدل‌ها در بازسازی تغییرات آن دقت کمتری دارند.

---

1. Normal Root Mean Squared Error
2. Kling-Gupta Efficiency

جدول ۴. مدل‌های مورد استفاده برای بازسازی داده‌های گمشده در این پژوهش.

ایستگاه	شاخص ارزیابی	محدوده شاخص	مدل‌های مورد استفاده			
			MICE	MICE-GBR	MICE-XGB	MICE-LGBM
بناب	R^2	بیشینه (متغیر)	(Tmin) ۰/۹۸۹	(Tmean) ۰/۹۹۲	(Tmean) ۰/۹۹۳	(Tmean) ۰/۹۹۲
		کمینه (متغیر)	(SLPmean) ۰/۰۲۶	(SLPmean) ۰/۴۳۹	(SLPmean) ۰/۴۴۵	(SLPmean) ۰/۴۲۹
	NRMSE	بیشینه (متغیر)	(CLDmean) ۴۵/۷۲۷	(CLDmean) ۳۹/۹۸۳	(CLDmean) ۳۸/۵۹۴	(CLDmean) ۳۸/۴۹۵
		کمینه (متغیر)	(SLPmean) ۰/۴۴۷	(SLPmean) ۰/۳۳۹	(SLPmean) ۰/۳۴۱	(SLPmean) ۰/۳۴۲
	|PBIAS|	بیشینه (متغیر)	(CLDmax) ۰/۰۱۶	(ET) ۰/۰۱۶	(CLDmean) ۰/۰۱۵	(ET) ۰/۰۱۴
		کمینه (متغیر)	(Sun) ۰/۰۰۰	(SLPmean) ۰/۰۰۰	(SLPmean) ۰/۰۰۰	(SLPmean) ۰/۰۰۰
	KGE	بیشینه (متغیر)	(Tmean) ۰/۹۹۰	(Tmean) ۰/۹۹۱	(Tmean) ۰/۹۹۲	(Tmean) ۰/۹۸۹
		کمینه (متغیر)	(SLPmean) ۰/۴۲۳	(SLPmean) ۰/۵۵۶	(SLPmean) ۰/۵۷۷	(SLPmean) ۰/۵۶۵
مراغه	R^2	بیشینه (متغیر)	(Tmean) ۰/۹۸۸	(Tmean) ۰/۹۹۰	(Tmean) ۰/۹۹۴	(Tmean) ۰/۹۹۱
		کمینه (متغیر)	(SLPmean) ۰/۳۸۷	(SLPmean) ۰/۴۸۶	(SLPmean) ۰/۴۸۶	(SLPmean) ۰/۵۰۳
	NRMSE	بیشینه (متغیر)	(CLDmean) ۴۰/۵۶۰	(CLDmean) ۳۷/۵۰۴	(CLDmean) ۳۹/۲۹۹	(CLDmean) ۳۸/۲۷۶
		کمینه (متغیر)	(SLPmean) ۰/۶۴۳	(SLPmean) ۰/۵۸۹	(SLPmean) ۰/۵۸۹	(SLPmean) ۰/۵۷۹
	|PBIAS|	بیشینه (متغیر)	(ET) ۰/۰۰۸	(CLDmean) ۰/۰۱۳	(ET) ۰/۰۰۷	(CLDmax) ۰/۰۱۳
		کمینه (متغیر)	(SLPmean) ۰/۰۰۰	(SLPmean) ۰/۰۰۰	(SLPmean) ۰/۰۰۰	(SLPmean) ۰/۰۰۰
	KGE	بیشینه (متغیر)	(RHmean) ۰/۹۸۱	(Tmean) ۰/۹۹۰	(Tmean) ۰/۹۹۵	(Tmax) ۰/۹۹۳
		کمینه (متغیر)	(SLPmean) ۰/۵۵۵	(SLPmean) ۰/۶۵۱	(SLPmean) ۰/۶۶۸	(SLPmean) ۰/۶۷۳
سقز	R^2	بیشینه (متغیر)	(Tmean) ۰/۹۷۴	(Tmean) ۰/۹۸۶	(Tmean) ۰/۹۸۶	(Tmean) ۰/۹۸۴
		کمینه (متغیر)	(SLPmean) ۰/۴۲۳	(SLPmean) ۰/۵۱۳	(SLPmean) ۰/۵۴۱	(SLPmean) ۰/۵۴۲
	NRMSE	بیشینه (متغیر)	(CLDmean) ۴۴/۲۷۶	(CLDmean) ۳۸/۱۹۳	(CLDmean) ۳۹/۶۸۱	(CLDmean) ۴۲/۰۳۴
		کمینه (متغیر)	(SLPmean) ۰/۳۱۲	(SLPmean) ۰/۲۸۷	(SLPmean) ۰/۲۷۸	(SLPmean) ۰/۲۸۰
	|PBIAS|	بیشینه (متغیر)	(CLDmax) ۰/۰۲۳	(CLDmean) ۰/۰۱۶	(CLDmax) ۰/۰۱۵	(RHmax) ۰/۰۰۹
		کمینه (متغیر)	(SLPmean) ۰/۰۰۰	(SLPmean) ۰/۰۰۰	(SLPmean) ۰/۰۰۰	(SLPmean) ۰/۰۰۰
	KGE	بیشینه (متغیر)	(RHmean) ۰/۹۸۳	(RHmean) ۰/۹۸۴	(RHmean) ۰/۹۸۶	(RHmean) ۰/۹۸۵
		کمینه (متغیر)	(SLPmean) ۰/۵۲۶	(SLPmean) ۰/۶۲۲	(SLPmean) ۰/۶۴۸	(SLPmean) ۰/۶۵۵
سراب	R^2	بیشینه (متغیر)	(Tmean) ۰/۹۷۷	(Tmean) ۰/۹۸۸	(Tmean) ۰/۹۸۹	(Tmean) ۰/۹۸۷
		کمینه (متغیر)	(SLPmean) ۰/۲۷۸	(SLPmean) ۰/۳۳۲	(SLPmean) ۰/۳۲۸	(SLPmean) ۰/۳۲۳
	NRMSE	بیشینه (متغیر)	(CLDmean) ۳۴/۷۵۵	(CLDmean) ۳۲/۳۰۱	(CLDmean) ۳۲/۵۸۳	(CLDmean) ۳۲/۵۲۸
		کمینه (متغیر)	(SLPmean) ۰/۳۲۷	(SLPmean) ۰/۳۱۴	(SLPmean) ۰/۳۱۵	(SLPmean) ۰/۳۱۶
	|PBIAS|	بیشینه (متغیر)	(CLDmax) ۰/۰۱۱	(CLDmax) ۰/۰۰۷	(CLDmax) ۰/۰۱۶	(CLDmean) ۰/۰۱۷
		کمینه (متغیر)	(RHmax) ۰/۰۰۰	(SLPmean) ۰/۰۰۰	(SLPmean) ۰/۰۰۰	(Tmean) ۰/۰۰۰
	KGE	بیشینه (متغیر)	(VPmean) ۰/۹۷۶	(Tmean) ۰/۹۸۹	(Tmean) ۰/۹۹۲	(TDmean) ۰/۹۹۱
		کمینه (متغیر)	(SLPmean) ۰/۳۷۵	(SLPmean) ۰/۴۶۳	(SLPmean) ۰/۴۷۸	(SLPmean) ۰/۴۸۲
تبریز	R^2	بیشینه (متغیر)	(Tmean) ۰/۹۸۸	(Tmean) ۰/۹۹۳	(Tmean) ۰/۹۹۵	(Tmean) ۰/۹۹۳
		کمینه (متغیر)	(SLPmean) ۰/۳۹۶	(SLPmean) ۰/۵۲۳	(SLPmean) ۰/۵۲۲	(SLPmean) ۰/۵۳۰
	NRMSE	بیشینه (متغیر)	(CLDmean) ۳۱/۷۷۳	(CLDmean) ۳۰/۰۸۵	(CLDmean) ۳۰/۱۹۶	(CLDmean) ۳۰/۰۰۰
		کمینه (متغیر)	(SLPmean) ۰/۳۴۰	(SLPmean) ۰/۳۰۲	(SLPmean) ۰/۳۰۲	(SLPmean) ۰/۲۹۹
	|PBIAS|	بیشینه (متغیر)	(TDmean) ۰/۰۰۷	(CLDmean) ۰/۰۱۱	(Sun) ۰/۰۰۵۶	(CLDmean) ۰/۰۱۶
		کمینه (متغیر)	(SLPmean) ۰/۰۰۰	(Tmean) ۰/۰۰۰	(SLPmean) ۰/۰۰۰	(Tmean) ۰/۰۰۰
	KGE	بیشینه (متغیر)	(Tmax) ۰/۹۷۵	(Tmean) ۰/۹۹۴	(Tmean) ۰/۹۹۵	(TDmean) ۰/۹۹۲
		کمینه (متغیر)	(SLPmean) ۰/۴۹۴	(SLPmean) ۰/۶۱۴	(SLPmean) ۰/۶۳۳	(SLPmean) ۰/۶۲۷
ارومیه	R^2	بیشینه (متغیر)	(Tmean) ۰/۹۸۵	(Tmean) ۰/۹۹۶	(Tmean) ۰/۹۹۶	(Tmean) ۰/۹۹۶
		کمینه (متغیر)	(SLPmean) ۰/۳۲۷	(SLPmean) ۰/۴۷۸	(SLPmean) ۰/۴۸۰	(WDmax) ۰/۴۶۸
	NRMSE	بیشینه (متغیر)	(CLDmean) ۳۶/۹۶۵	(CLDmean) ۳۰/۰۱۶	(CLDmean) ۲۹/۹۴۲	(CLDmean) ۲۹/۶۷۶
		کمینه (متغیر)	(SLPmean) ۰/۴۱۳	(SLPmean) ۰/۳۶۳	(SLPmean) ۰/۳۶۵	(SLPmean) ۰/۳۶۷
	|PBIAS|	بیشینه (متغیر)	(CLDmax) ۰/۰۱۷	(CLDmean) ۰/۰۱۹	(Tmin) ۰/۰۱۶	(CLDmean) ۰/۰۲۰
		کمینه (متغیر)	(SLPmean) ۰/۰۰۰	(SLPmean) ۰/۰۰۰	(SLPmean) ۰/۰۰۰	(SLPmean) ۰/۰۰۰
	KGE	بیشینه (متغیر)	(VPmean) ۰/۹۷۹	(SVPmean) ۰/۹۹۵	(SVPmean) ۰/۹۹۶	(SVPmean) ۰/۹۹۴
		کمینه (متغیر)	(SLPmean) ۰/۴۲۳	(SLPmean) ۰/۵۶۴	(SLPmean) ۰/۵۷۹	(SLPmean) ۰/۵۶۷

بیشترین مقدار NRMSE (۴۵/۷۲۷) در ایستگاه بناب برای متغیر میانگین ابرناکی (CLDmean) و با مدل MICE ثبت شد، در حالی‌که کمترین مقدار آن (۰/۲۷۸) در ایستگاه سقز برای متغیر میانگین فشار سطح دریا (SLPmean) و با مدل MICE–XGB به‌دست آمد. به‌طور کلی، در مدل MICE مقادیر NRMSE به‌طور محسوسی بیشتر از سه مدل دیگر بوده است که بیانگر ضعف نسبی روش پایه در بازسازی دقیق داده‌ها نسبت به مدل‌های ترکیبی تقویتی است. همچنین، بیشترین مقادیر NRMSE در متغیر میانگین ابرناکی روزانه (CLDmean) و کمترین مقدار آن در متغیر میانگین فشار سطح دریا (SLPmean) مشاهده شد. لازم به ذکر است که اگرچه مقیاس عددی متغیرها می‌تواند بر مقدار خام RMSE تأثیرگذار باشد، اما از آن‌جا که در این پژوهش از شاخص نرمال‌شده (NRMSE) استفاده شده است، تفاوت‌های مشاهده‌شده عمدتاً ناشی از رفتار ذاتی متغیرهاست. درواقع، نوسانات زیاد و ماهیت ناپیوسته متغیر ابرناکی موجب افزایش خطا در بازسازی آن شده، در حالی‌که پایداری زمانی متغیر فشار سطح دریا باعث کاهش مقدار NRMSE در این متغیر گردیده است.

مقادیر شاخص |PBIAS| در تمامی ایستگاه‌ها، متغیرها و مدل‌های مورد بررسی کمتر از ۰/۰۲۵ درصد برآورد شد که بیانگر عملکرد فوق‌العاده دقیق مدل‌ها در بازسازی داده‌ها و نبود انحراف سیستماتیک میان داده‌های بازسازی‌شده و مشاهدات واقعی است. بیشترین مقدار |PBIAS| در ایستگاه سقز، مربوط به متغیر بیشینه ابرناکی روزانه (CLDmax) و با مدل MICE مشاهده شد. در مقابل، کمترین مقدار آن در متغیر میانگین فشار سطح دریا (SLPmean) به‌دست آمد. به‌طور کلی، بیشینه مقادیر |PBIAS| معمولاً در متغیرهای ابرناکی و تبخیر و تعرق (ET) رخ داده است، در حالی‌که کمترین مقادیر در متغیرهای فشاری ثبت شده‌اند. دلیل این الگو آن است که متغیرهای ابرناکی و تبخیر و تعرق به‌شدت تحت تأثیر نوسانات کوتاه‌مدت جوی، شرایط تابشی و رطوبتی موضعی قرار دارند و ماهیت غیرخطی و پراکنده‌تری نسبت به سایر متغیرها دارند؛ بنابراین مدل‌ها ممکن است در بازسازی آن‌ها دچار اندکی بایاس شوند. در مقابل، متغیر فشار سطح دریا از پایداری زمانی بالاتری برخوردار بوده و همبستگی قوی‌تری با سایر پارامترهای ترمودینامیکی دارد؛ به همین دلیل، خطای بایاس در بازسازی آن بسیار ناچیز است.

بر اساس نتایج جدول ۴، بیشترین مقدار شاخص KGE در ایستگاه ارومیه (۰/۹۹۶) برای متغیر میانگین فشار بخار اشباع (SVPmean) و در ایستگاه‌های مراغه و تبریز (۰/۹۹۵) برای متغیر میانگین دمای روزانه (Tmean) و هر سه مورد با مدل MICE–XGB به‌دست آمده است. به‌طور کلی، مقدار KGE در مدل‌های ترکیبی MICE با روش‌های تقویتی یادگیری ماشین به‌مراتب بالاتر از مدل پایه MICE بوده است که بیانگر هم‌خوانی بیشتر بین داده‌های بازسازی‌شده و مشاهدات واقعی در این مدل‌هاست. بیشترین مقادیر KGE معمولاً در متغیرهای دمایی، رطوبت نسبی و فشار بخار آب مشاهده شد، در حالی‌که کمینه مقدار آن مربوط به متغیر میانگین فشار سطح دریا (SLPmean) بوده است. علت این رفتار آن است که متغیرهای دمایی و رطوبتی دارای ساختار همبستگی قوی با یکدیگر و نوسانات منظم‌تری در بازه‌های زمانی روزانه هستند، بنابراین مدل‌های یادگیری تقویتی قادر به شناسایی بهتر الگوهای درونی آن‌ها بوده‌اند. در مقابل، فشار سطح دریا تحت تأثیر تغییرات دینامیکی ترازهای بالاتر جو و شرایط سینوپتیکی منطقه قرار دارد که منجر به افزایش خطا و کاهش همبستگی بین مقادیر بازسازی‌شده و واقعی در این متغیر می‌شود.

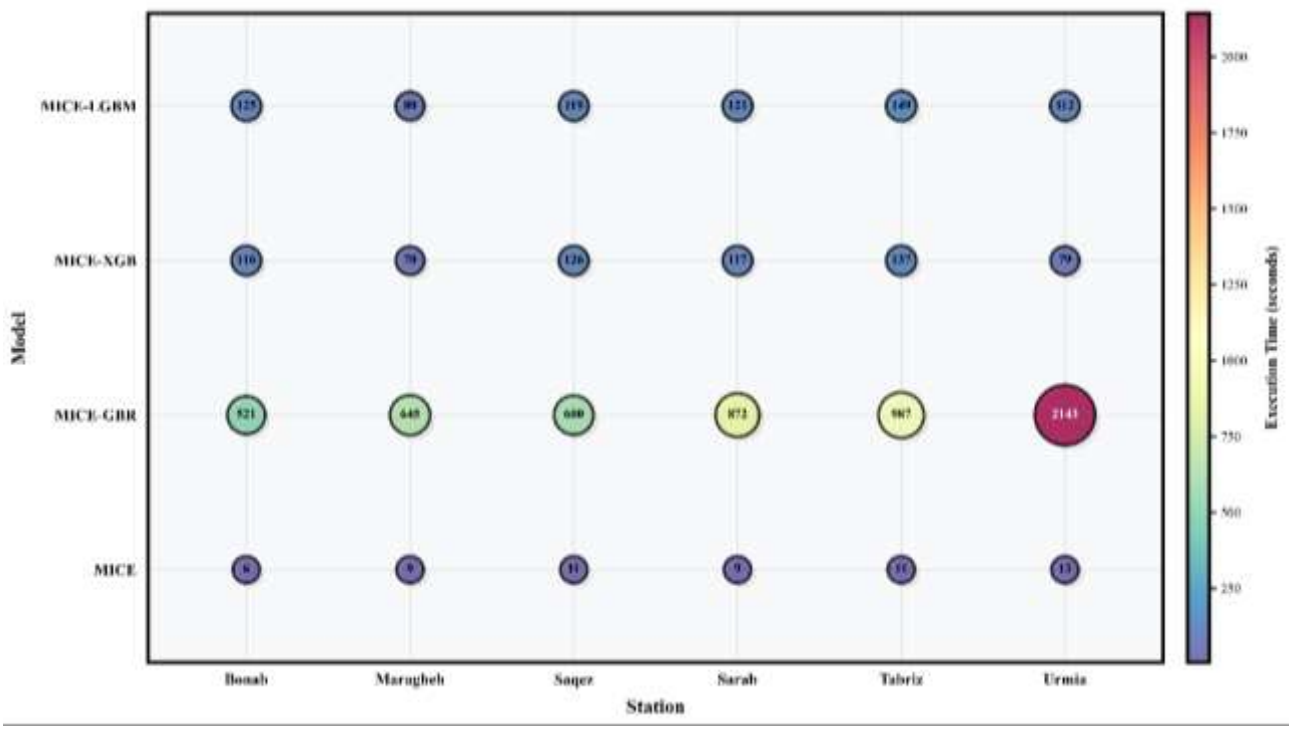
در مجموع، نتایج شاخص‌های آماری در تمامی ایستگاه‌ها نشان داد که مدل‌های ترکیبی مبتنی بر رویکردهای تقویتی یادگیری ماشین (به‌ویژه MICE–XGB و MICE–LGBM) عملکرد بسیار بهتری نسبت به مدل پایه MICE در بازسازی داده‌های گمشده اقلیمی دارند. این مدل‌ها ضمن افزایش ضریب تعیین (R^2) و کارایی کلینگ–گوپتا (KGE)، مقادیر خطا (NRMSE) و بایاس نسبی (|PBIAS|) را به حداقل رسانده‌اند. همچنین، دقت مدل‌ها در متغیرهای دمایی، رطوبتی و فشار بخار آب به‌مراتب بالاتر از متغیرهای ابرناکی و تبخیر و تعرق بوده است که نشان از حساسیت بیشتر متغیرهای اخیر به نوسانات کوتاه‌مدت جوی دارد.

علاوه بر ارزیابی دقت، زمان اجرای هر مدل نیز در بخش بعدی مورد بررسی قرار گرفت تا تعادلی میان دقت بازسازی و کارایی محاسباتی مدل‌ها تحلیل شود.

تحلیل زمان اجرای مدل‌ها

تمامی محاسبات این پژوهش بر روی یک سیستم محاسباتی یکسان با پردازنده Intel Core i7 (نسل یازدهم) با فرکانس پایه ۲٫۸ GHz، حافظه RAM برابر با ۱۲ گیگابایت و سیستم‌عامل Windows 10 (64-bit) انجام شد. پیاده‌سازی مدل‌ها در محیط Python نسخه ۳٫۱۰ صورت گرفت و از کتابخانه‌های scikit-learn، xgboost و lightgbm استفاده گردید. زمان اجرای هر مدل با استفاده از ماژول time در پایتون و تحت شرایط سخت‌افزاری و نرم‌افزاری یکسان اندازه‌گیری شد تا امکان مقایسه منصفانه و تکرارپذیری نتایج فراهم شود.

به‌منظور ارزیابی جنبه محاسباتی مدل‌های مورد استفاده در بازسازی داده‌های گمشده اقلیمی، زمان اجرای چهار مدل MICE، MICE–GBR، MICE–XGB و MICE–LGBM در شش ایستگاه منتخب حوضه آبریز دریاچه ارومیه محاسبه و مقایسه گردید. این تحلیل با هدف بررسی کارایی زمانی مدل‌ها و تعیین بهینه‌ترین روش از نظر نسبت دقت به هزینه محاسباتی انجام شد. نتایج حاصل از زمان اجرای مدل‌ها در قالب نمودار حبابی (شکل ۲) نشان داد که تفاوت قابل‌توجهی در مدت‌زمان اجرای روش‌های مختلف وجود دارد، به‌طوری که مدل پایه MICE کمترین زمان و مدل‌های تقویتی به‌ویژه MICE–GBR بیشترین زمان محاسباتی را به خود اختصاص داده‌اند.



شکل ۲. مقایسه زمان اجرای مدل‌های مختلف بازسازی داده در ایستگاه‌های منتخب با استفاده از نمودار حبابی (اندازه حباب‌ها بیانگر میزان زمان محاسباتی مدل‌ها (بر حسب ثانیه) در هر ایستگاه است.)

مدل MICE به‌دلیل ماهیت ساده و خطی خود، کمترین زمان اجرا را در میان تمامی مدل‌ها داشته است. زمان اجرای این مدل در ایستگاه‌های مختلف بین ۶ تا ۱۳ ثانیه متغیر بوده و در مقایسه با مدل‌های ترکیبی، چندین برابر سریع‌تر اجرا شده است. با این حال، نتایج پیشین نشان داد که اگرچه سرعت این مدل بالا است، اما در بازسازی دقیق داده‌ها به‌ویژه برای متغیرهای دارای رفتار غیرخطی (مانند ابرناکی یا تبخیر و تعرق)، از عملکرد ضعیف‌تری برخوردار است. در مقابل، مدل MICE–GBR بیشترین زمان محاسباتی را در بین تمام روش‌ها نشان داد؛ به‌ویژه در ایستگاه ارومیه که زمان اجرای آن به حدود ۲۱۴۳ ثانیه رسید. این امر به‌دلیل پیچیدگی ذاتی الگوریتم Gradient Boosting Regression است که شامل فرایند تکراری ساخت مجموعه‌ای از درخت‌های تصمیم ضعیف و به‌روزرسانی تدریجی وزن‌های خطا در هر مرحله است. این ویژگی گرچه باعث افزایش دقت مدل می‌شود، اما هزینه‌ی محاسباتی بالایی دارد و اجرای آن در مجموعه‌داده‌های بزرگ‌تر یا دارای تعداد متغیر زیاد می‌تواند زمان‌بر باشد.

مدل‌های MICE–XGB و MICE–LGBM با به‌کارگیری ساختارهای بهینه‌سازی‌شده و استفاده از روش‌های موازی‌سازی (Parallel Processing) و تکنیک‌های کاهش پیچیدگی محاسباتی، عملکرد بسیار کارآمدتری از خود نشان دادند. میانگین زمان اجرای مدل MICE–XGB در بین ایستگاه‌ها حدود ۱۰۸ ثانیه و برای مدل MICE–LGBM حدود ۱۱۸ ثانیه برآورد شد. این دو مدل در بیشتر ایستگاه‌ها زمان اجرای مشابهی داشتند، اما در عین حال دقت بالاتری نسبت به MICE و سرعت بهتری نسبت به MICE–GBR ارائه کردند. این ویژگی ناشی از طراحی الگوریتمی پیشرفته‌ی آن‌هاست؛ به‌گونه‌ای که XGBoost از تکنیک‌های Gradient Pruning و Tree Regularization برای کاهش پیچیدگی استفاده می‌کند، در حالی که LightGBM با بهره‌گیری از Histogram-based Decision Tree و Leaf-wise Growth Strategy قادر است زمان آموزش را به شکل چشمگیری کاهش دهد.

در مقایسه بین ایستگاه‌ها، مشاهده شد که زمان اجرای مدل‌ها در ایستگاه ارومیه بیشترین مقدار را دارد. علت این موضوع را می‌توان به حجم بالاتر داده‌ها، تعداد بیشتر مقادیر گمشده و تنوع اقلیمی بالاتر این ایستگاه نسبت داد. در مقابل، ایستگاه بناب با کمترین حجم داده و پراکندگی زمانی کمتر، کمترین زمان اجرا را برای تمام مدل‌ها نشان داد.

به‌طور کلی، نتایج این بخش بیانگر آن است که استفاده از مدل‌های ترکیبی تقویتی منجر به افزایش زمان محاسباتی نسبت به روش پایه MICE می‌شود، اما این افزایش در برابر بهبود چشمگیر دقت مدل‌ها قابل توجیه است. در این میان، مدل MICE–XGB به‌عنوان بهینه‌ترین روش از نظر توازن میان دقت و کارایی زمانی شناسایی شد؛ زیرا ضمن حفظ دقت بالا در شاخص‌های ارزیابی (R^2، KGE و NRMSE پایین)، زمان محاسباتی نسبتاً کمی را نیز نیاز دارد. بنابراین، می‌توان نتیجه گرفت که مدل‌های تقویتی مدرن مانند XGBoost و LightGBM با ساختارهای سبک‌تر و یادگیری کارآمدتر، گزینه‌های مناسب‌تری برای بازسازی داده‌های اقلیمی در مقیاس‌های بزرگ‌تر هستند.

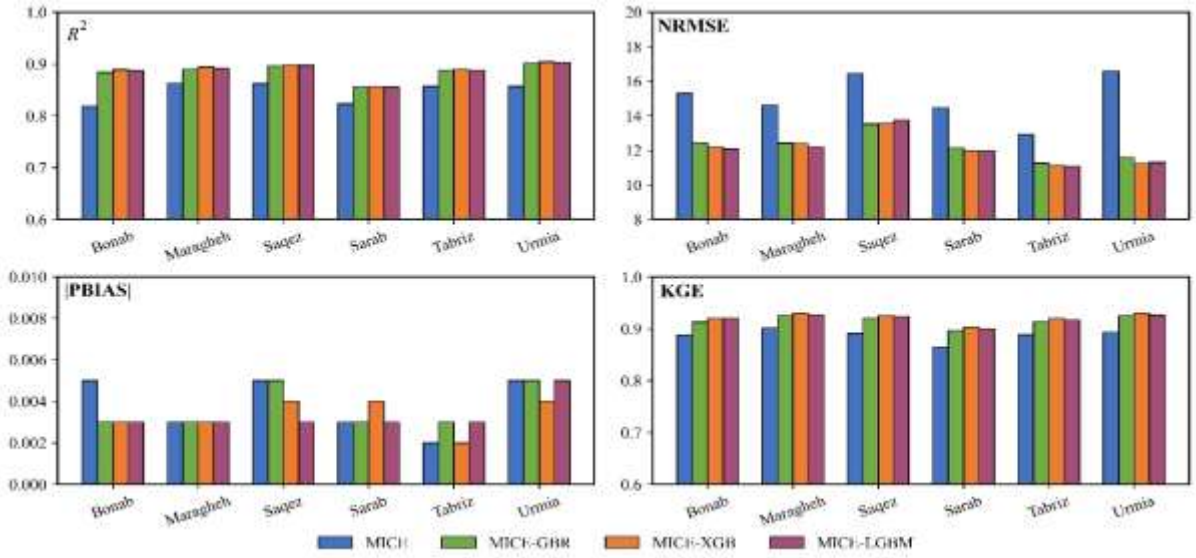
در ادامه، برای درک بهتر رفتار مدل‌ها در شرایط مکانی و اقلیمی متفاوت، عملکرد هر یک از مدل‌ها در سطح ایستگاه‌های منتخب و همچنین در متغیرهای اقلیمی مختلف به‌صورت تفصیلی مورد تحلیل و مقایسه قرار خواهد گرفت.

تحلیل عملکرد مدل‌ها در سطح ایستگاه‌های منتخب

برای بررسی دقیق‌تر عملکرد مدل‌ها در مقیاس مکانی، میانگین شاخص‌های آماری R^2، NRMSE، |PBIAS| و KGE درشش ایستگاه منتخب حوضه آبریز دریاچه ارومیه محاسبه و در قالب نمودارهای ستونی مقایسه شدند (شکل ۳). بررسی نمودارها نشان داد که مدل‌های ترکیبی مبتنی بر یادگیری تقویتی (به‌ویژه MICE–XGB و MICE–LGBM) در تمامی ایستگاه‌ها عملکرد دقیق‌تر و پایاتری نسبت به مدل پایه MICE دارند.

نتایج نشان دادند که در تمامی ایستگاه‌ها، مقدار R^2 در مدل‌های ترکیبی بالاتر از مدل پایه MICE است. بیشترین مقدار R^2 به مدل MICE–XGB در ایستگاه ارومیه (۰/۹۰۴) و کمترین مقدار به مدل MICE در ایستگاه بناب (۰/۸۱۹) تعلق دارد. مدل‌های MICE–XGB و MICE–LGBM در بیشتر ایستگاه‌ها مقدار R^2 بالاتر از ۰/۸۹ ثبت کرده‌اند که بیانگر همبستگی بسیار قوی بین داده‌های بازسازی‌شده و مقادیر مشاهداتی است. تفاوت اندک بین MICE–XGB و MICE–LGBM نشان‌دهنده پایداری و هم‌گرایی نتایج این دو مدل است.

تحلیل مقادیر NRMSE نشان داد که با افزایش پیچیدگی و هوشمندی مدل‌ها، مقدار خطای نرمال‌شده به‌طور یکنواخت کاهش یافته است. بیشترین مقدار NRMSE مربوط به مدل MICE در ایستگاه ارومیه (۱۶/۵۹۴) و کمترین مقدار آن مربوط به مدل MICE–XGB در همان ایستگاه (۱۱/۲۳۰) بود. میانگین NRMSE مدل MICE–XGB در تمام ایستگاه‌ها حدود ۱۲/۷ درصد کمتر از مدل پایه MICE بوده است که نشان از بهبود چشمگیر دقت در بازسازی دارد. کاهش خطا در مدل‌های تقویتی به‌ویژه در ایستگاه‌های دارای شرایط اقلیمی متنوع‌تر (مانند تبریز و ارومیه) بیشتر مشهود است که نشان‌دهنده توانایی بالاتر این مدل‌ها در شناسایی روابط غیرخطی میان متغیرهاست.



شکل ۳. شاخص‌های ارزیابی عملکرد مدل‌های بازسازی داده‌های گمشده اقلیمی در ایستگاه‌های منتخب حوضه دریاچه ارومیه

در تمامی مدل‌ها، مقدار |PBIAS| بسیار پایین (کمتر از ۰/۰۰۵) بوده که نشان‌دهنده عدم وجود انحراف سیستماتیک بین داده‌های بازسازی‌شده و داده‌های واقعی است. با این حال، مدل‌های MICE–XGB و MICE–LGBM مقادیر بایاس اندکی کمتر نسبت به MICE و MICE–GBR نشان دادند، که بیانگر بازسازی متعادل‌تر داده‌ها و حفظ میانگین آماری متغیرها در این دو مدل است. کمترین مقدار بایاس مطلق در ایستگاه تبریز برای مدل MICE–XGB (۰/۰۰۲) ثبت شد.

نتایج شاخص کارایی کلینگ–گوپتا (KGE) نیز به‌خوبی برتری مدل‌های تقویتی را تأیید کرد. بیشترین مقدار KGE در ایستگاه‌های مراغه و ارومیه برای مدل MICE–XGB (۰/۹۳) و کمترین مقدار آن در ایستگاه سراب برای مدل MICE (۰/۸۶۵) مشاهده شد. افزایش پیوسته مقادیر KGE از MICE به سمت MICE–GBR، MICE–XGB و MICE–LGBM روندی کاملاً مشخص داشت و این الگو در تمامی ایستگاه‌ها یکسان بود. مقادیر بالای KGE (بیش از ۰/۹۰) در مدل‌های تقویتی نشان می‌دهد که این مدل‌ها نه‌تنها در بازسازی روند داده‌ها دقیق‌تر عمل کرده‌اند، بلکه در حفظ نوسانات طبیعی و الگوهای درونی داده‌های اقلیمی نیز موفق بوده‌اند.

تحلیل‌های مکانی نشان داد که عملکرد مدل‌ها در ایستگاه‌های پایین‌ارتفاع و اقلیم معتدل‌تر مانند ارومیه، بناب و تبریز بهتر بود. این امر با ویژگی‌های فیزیکی–اقلیمی این ایستگاه‌ها مرتبط است: نوسانات روزانه دما، رطوبت و تبخیر نسبی در این مناطق کمتر و پایدارتر بوده و رفتار متغیرهای اقلیمی نزدیک به روندهای میانگین بلندمدت است، که بازسازی داده‌ها را ساده‌تر و دقیق‌تر می‌کند. در مقابل، ایستگاه‌های مرتفع‌تر و سردتر مانند سراب و سقز دارای نوسانات شدیدتر روزانه و پراکندگی بالاتر داده‌ها هستند. تغییرات دمای روزانه بیشتر، افزایش تبخیر و رطوبت متغیر و رفتار غیرخطی شدیدتر متغیرهای اقلیمی، موجب افزایش چالش در پیش‌بینی مقادیر گمشده شده و خطای بازسازی در این ایستگاه‌ها کمی بالاتر است. با این حال، حتی در این ایستگاه‌ها نیز مدل‌های تقویتی (به‌ویژه MICE–XGB و MICE–LGBM) توانسته‌اند دقتی بالا و خطایی بسیار اندک را حفظ کنند که نشان‌دهنده‌ی پایداری مکانی و استحکام عمومی این مدل‌ها در بازسازی داده‌های اقلیمی حوضه دریاچه ارومیه است. لازم به ذکر است که داده‌ها برای هر ایستگاه به‌صورت مستقل ارزیابی و مدل‌ها به صورت جداگانه آموزش داده شدند، به طوری که نتایج هر ایستگاه منعکس‌کننده شرایط خاص اقلیمی و مکانی آن ایستگاه باشد. این رویکرد امکان ارزیابی مستقل توانایی مدل‌ها در بازسازی داده‌ها در شرایط اقلیمی متنوع را فراهم کرد.

در مجموع، نتایج این بخش نشان می‌دهد که مدل‌های ترکیبی مبتنی بر یادگیری تقویتی، علاوه بر دقت بالا در شاخص‌های آماری، از پایداری مکانی مطلوبی نیز برخوردارند. در بخش‌های بعدی، عملکرد این مدل‌ها در سطح ایستگاه‌های منتخب و در بازسازی متغیرهای اقلیمی مختلف به‌صورت دقیق‌تر مورد تحلیل قرار خواهد گرفت.

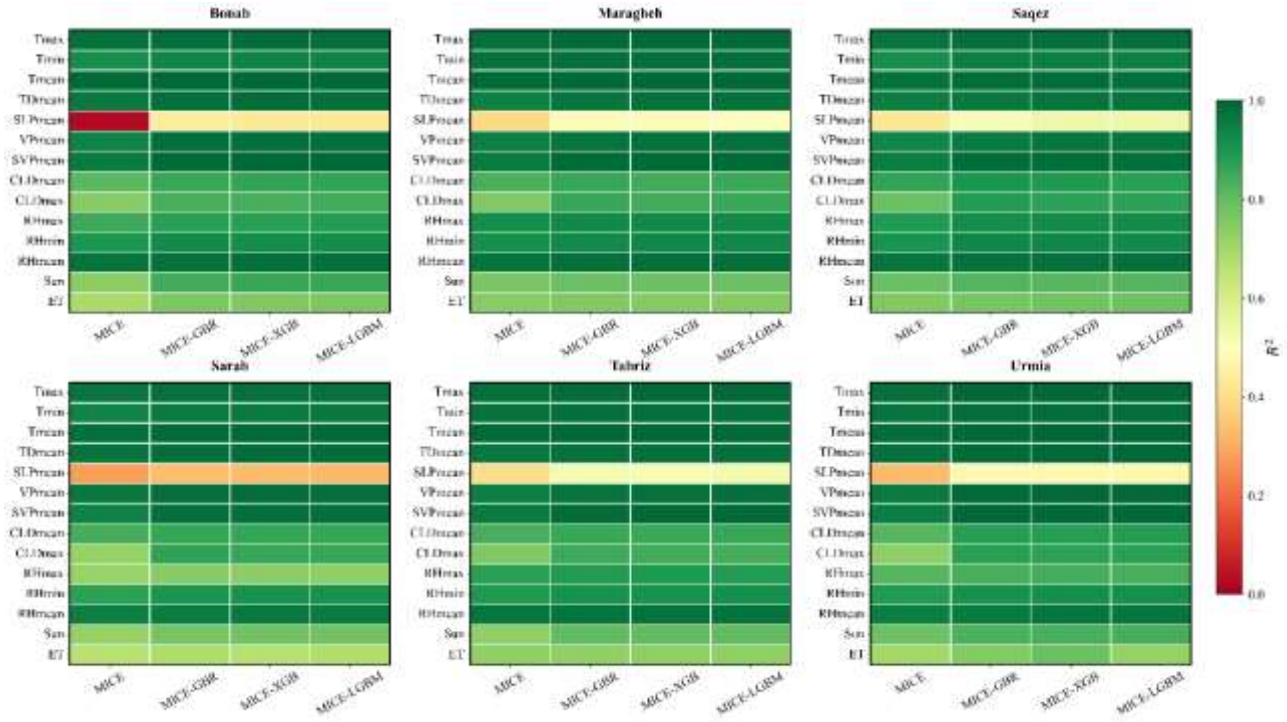
ارزیابی عملکرد مدل‌ها در متغیرهای اقلیمی مختلف

در این بخش، به‌منظور تحلیل دقیق‌تر رفتار مدل‌ها، عملکرد آن‌ها در بازسازی ۱۴ متغیر مختلف اقلیمی شامل دما، فشار، رطوبت، ابرناکی، تابش و تبخیر و تعرق با استفاده از سه شاخص آماری R^2، NRMSE و KGE مورد بررسی قرار گرفت (شکل‌های ۴، ۵ و ۶). برای هر شاخص، مقادیر به‌صورت نقشه حرارتی نمایش داده شد تا تفاوت عملکرد مدل‌ها در میان متغیرها و ایستگاه‌ها به‌صورت بصری و مقایسه‌پذیر مشخص شود.

تحلیل بر اساس شاخص R^2

به‌منظور بررسی میزان دقت مدل‌ها در بازسازی داده‌های مفقود، شاخص تعیین (R^2) برای تمامی متغیرهای اقلیمی و در شش ایستگاه منتخب محاسبه و در قالب نقشه‌ی حرارتی ترسیم شد (شکل ۴).

بر اساس نتایج، مقدار R^2 در تمامی مدل‌ها و متغیرها عموماً بیش از ۰/۸۰ بوده و این امر بیانگر همبستگی بسیار قوی بین داده‌های بازسازی‌شده و مقادیر مشاهداتی واقعی است. در میان مدل‌های مورد بررسی، مدل‌های ترکیبی مبتنی بر روش‌های تقویتی (به‌ویژه MICE–XGB و MICE–LGBM) عملکرد بهتری نسبت به مدل پایه MICE داشته‌اند، به‌گونه‌ای که در اغلب ایستگاه‌ها بیشترین مقدار R^2 مربوط به این دو مدل بوده است. از نظر فضایی، مقادیر R^2 در ایستگاه‌های مرکزی و شمالی حوضه (تبریز، مراغه و ارومیه) بالاتر از سایر نقاط برآورد شد که می‌تواند ناشی از پایداری بیشتر داده‌ها، همبستگی مکانی قوی‌تر و تراکم ایستگاه‌های مجاور باشد. در مقابل، در ایستگاه‌های جنوبی (به‌ویژه سقز) اندکی کاهش در مقدار R^2 مشاهده گردید که احتمالاً به دلیل نوسانات محلی و شرایط توپوگرافی متغیر منطقه است.



شکل ۴- نقشه حرارتی شاخص R^2 برای ارزیابی عملکرد مدل‌های مختلف در بازسازی داده‌های اقلیمی در ایستگاه‌های منتخب حوضه آبریز دریاچه ارومیه.

از نظر متغیرهای اقلیمی، بیشترین مقدار R^2 معمولاً در متغیرهای دمایی (به‌ویژه میانگین دمای روزانه Tmean و دمای بیشینه Tmax) مشاهده شد. این متغیرها از پایداری زمانی و الگوی تغییرات منظم‌تری برخوردارند و بنابراین مدل‌ها توانسته‌اند وابستگی‌های درونی آن‌ها را به‌خوبی شناسایی کنند. در مقابل، کمترین مقدار R^2 در بیشتر ایستگاه‌ها مربوط به متغیر میانگین فشار سطح دریا (SLPmean) بوده است. این متغیر به‌شدت تحت تأثیر نوسانات دینامیکی و سینوپتیکی جو است و نوسانات آن در مقیاس‌های زمانی کوتاه‌مدت کمتر قابل پیش‌بینی است؛ ازاین‌رو بازسازی آن با دقت کمتری همراه بوده است.

در مجموع، می‌توان نتیجه گرفت که مدل‌های تقویتی (به‌ویژه MICE–XGB) توانسته‌اند ساختار همبستگی میان متغیرها را بهتر درک کرده و داده‌های مفقود را با دقت بالاتری بازسازی کنند.

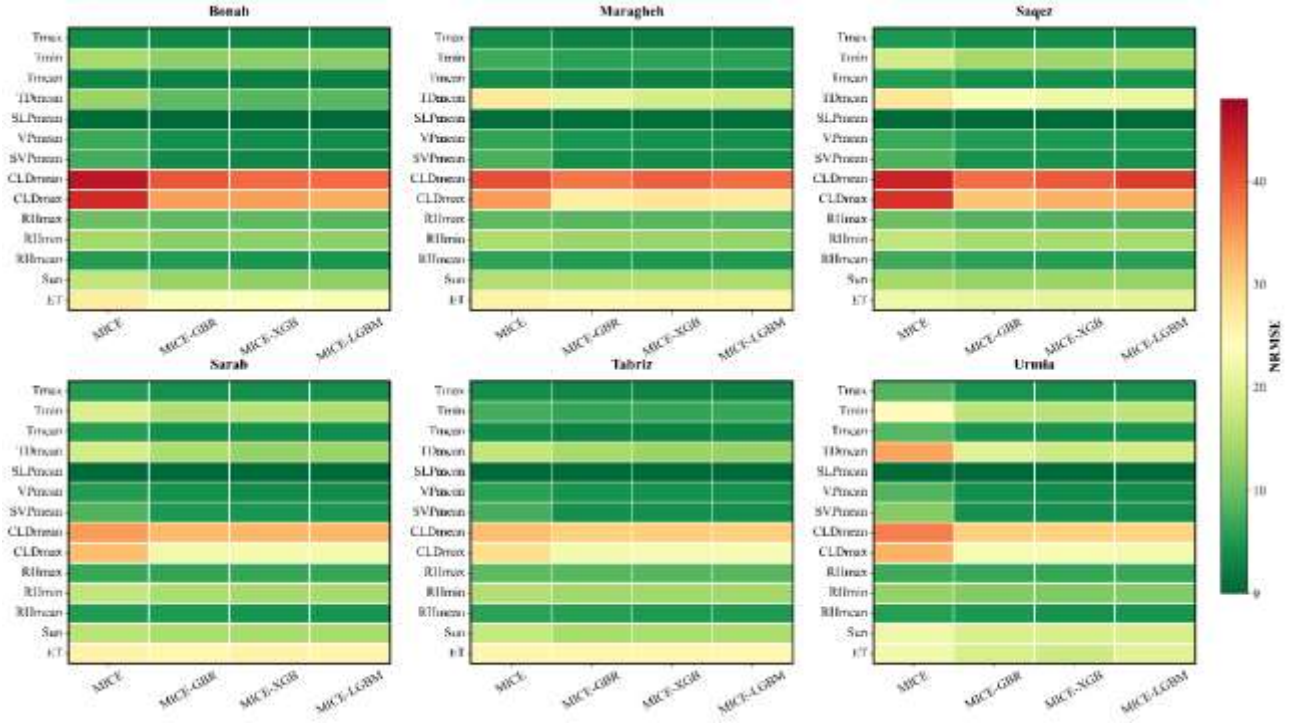
تحلیل بر اساس شاخص NRMSE

به‌منظور ارزیابی میزان دقت عددی مدل‌ها در بازسازی داده‌های مفقود، شاخص خطای نرمال‌شده ریشه میانگین مربعات (NRMSE)برای تمامی متغیرهای اقلیمی و ایستگاه‌ها محاسبه و در قالب نقشه‌ی حرارتی نمایش داده شد (شکل ۵).

بر اساس نتایج، مقدار NRMSE در اغلب ایستگاه‌ها کمتر از ۲۰ درصد بوده است که نشان‌دهنده دقت بسیار مطلوب مدل‌ها در بازسازی داده‌ها است. در بین روش‌ها، مدل پایه MICE بیشترین مقادیر خطا را به خود اختصاص داد، در حالی‌که مدل‌های ترکیبی مبتنی بر الگوریتم‌های تقویتی، به‌ویژه MICE–XGB و MICE–LGBM، کمترین خطا را در اغلب متغیرها نشان دادند. این امر نشان می‌دهد که ترکیب روش‌های تقویتی یادگیری ماشین با چارچوب چندتکاملی MICE موجب بهبود چشمگیر عملکرد مدل در بازسازی داده‌ها شده است. از نظر فضایی، در ایستگاه‌های جنوبی و مرتفع‌تر (مانند سقز و سراب) مقادیر NRMSE اندکی بیشتر مشاهده شد که می‌تواند ناشی از نوسانات بیشتر شرایط محلی، اثرات توپوگرافی و تغییرات روزانه شدیدتر متغیرهای اقلیمی در این نواحی باشد. در مقابل، ایستگاه‌های مرکزی و غربی حوضه (مانند تبریز و ارومیه) که از شرایط اقلیمی معتدل‌تر و پایدارتری برخوردارند، دارای مقادیر پایین‌تر خطا بودند.

از نظر نوع متغیر اقلیمی، بیشترین مقدار NRMSE معمولاً در متغیر میانگین ابرناکی (CLDmean) و در برخی موارد تبخیر و تعرق روزانه (ET) مشاهده شد. این متغیرها دارای رفتار ناپیوسته، پراکندگی زیاد و تأثیرپذیری بالا از شرایط موضعی مانند تابش و رطوبت لحظه‌ای هستند، که بازسازی دقیق آن‌ها را دشوار می‌سازد. در مقابل، کمترین مقادیر NRMSE مربوط به میانگین فشار سطح دریا (SLPmean) و میانگین دمای روزانه (Tmean) بوده است. پایداری زمانی بالا و رفتار منظم‌تر این متغیرها موجب شده مدل‌ها بتوانند

با دقت بیشتری الگوهای تغییر آن‌ها را بازسازی کنند.



شکل ۵- نقشه حرارتی شاخص NRMSE برای ارزیابی خطای نسبی مدل‌های مختلف در بازسازی داده‌های اقلیمی در ایستگاه‌های منتخب حوضه آبریز دریاچه ارومیه

به‌طور کلی، نتایج شاخص NRMSE کاملاً با یافته‌های شاخص R^2 هم‌خوانی دارد؛ یعنی هرجا R^2 افزایش یافته، مقدار NRMSE کاهش پیدا کرده است که این امر نشان‌دهنده پایداری و صحت بالای نتایج مدل‌ها در ارزیابی متقاطع شاخص‌ها است.

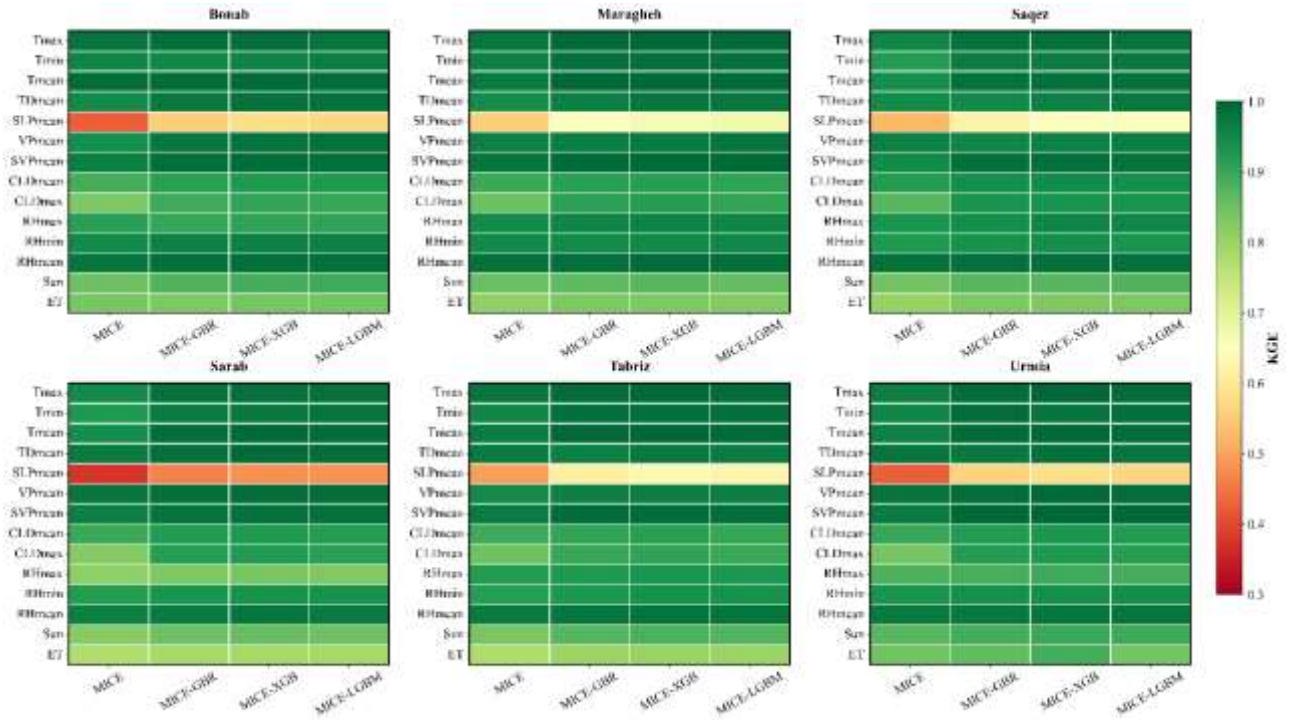
تحلیل بر اساس شاخص KGE

به‌منظور بررسی جامع‌تر میزان تطابق آماری داده‌های بازسازی‌شده با داده‌های واقعی، شاخص KGE برای تمامی متغیرهای اقلیمی و ایستگاه‌ها محاسبه و در قالب نقشه حرارتی ترسیم شد (شکل ۶). این شاخص به‌عنوان یکی از معیارهای پیشرفته ارزیابی مدل، هم‌زمان سه مؤلفه‌ی ضریب همبستگی، انحراف میانگین و نسبت انحراف معیار داده‌های بازسازی‌شده به مشاهدات را در نظر می‌گیرد و مقدار نزدیک به ۱ بیانگر دقت بسیار بالا و هم‌خوانی مطلوب مدل است.

بر اساس نتایج، مقادیر KGE در تمامی ایستگاه‌ها و مدل‌ها عموماً بیش از ۰/۸ و در بسیاری از موارد بیش از ۰/۹ برآورد شد که نشان‌دهنده همبستگی قوی بین داده‌های بازسازی‌شده و واقعی است. در بین مدل‌ها، MICE–XGB و MICE–LGBM بالاترین مقادیر KGE را در اغلب متغیرها و ایستگاه‌ها کسب کردند، که مؤید برتری مدل‌های تقویتی یادگیری ماشین نسبت به روش پایه MICE است. مدل MICE در مقایسه با مدل‌های ترکیبی، در برخی متغیرها مانند میانگین فشار سطح دریا (SLPmean) یا میانگین فشار بخار (VPmean) مقادیر نسبتاً پایین‌تری از KGE نشان داد که بیانگر توان کمتر آن در بازسازی الگوهای ناپیوسته و غیرخطی این متغیرها است.

از نظر فضایی، ایستگاه‌های تبریز و ارومیه دارای بیشترین مقادیر KGE بوده‌اند، در حالی‌که ایستگاه‌های سقز و سراب در برخی متغیرهای فشاری کمترین مقادیر را ثبت کردند. این امر می‌تواند ناشی از نوسانات سینوپتیکی و شرایط توپوگرافی پیچیده این مناطق باشد که سبب کاهش تطابق کامل بین داده‌های بازسازی‌شده و واقعی می‌گردد. از نظر نوع متغیر، بیشترین مقادیر KGE مربوط به متغیرهای دمایی (Tmean، Tmax، Tmin) و رطوبتی (RHmean، SVPmean) بوده است. این متغیرها دارای همبستگی درونی بالا، روندهای زمانی منظم و پایداری فصلی قابل توجهی هستند که منجر به افزایش دقت بازسازی می‌شود. در مقابل، کمترین مقادیر KGE برای میانگین فشار سطح دریا (SLPmean) ثبت شده است. این متغیر تحت تأثیر تغییرات دینامیکی ترازهای بالای جو، جابجایی

سامانه‌های فشار و تغییرات سینوپتیکی در مقیاس منطقه‌ای است، که منجر به افزایش خطا و کاهش همبستگی بازسازی می‌شود.



شکل ۶- نقشه حرارتی شاخص KGE برای ارزیابی خطای نسبی مدل‌های مختلف در بازسازی داده‌های اقلیمی در ایستگاه‌های منتخب حوضه آبریز دریاچه ارومیه

در مجموع، الگوی مشاهده‌شده در شاخص KGE هم‌راستا با شاخص‌های R^2 و NRMSE است و تأکید می‌کند که مدل‌های ترکیبی مبتنی بر تقویت یادگیری ماشین، به‌ویژه MICE–XGB، در بازسازی داده‌های اقلیمی حوضه آبریز دریاچه ارومیه عملکردی پایدار، دقیق و منسجم از خود نشان داده‌اند.

نتیجه‌گیری و پیشنهادها

نتایج این پژوهش نشان داد که استفاده از روش‌های ترکیبی مبتنی بر الگوریتم‌های تقویتی در چارچوب چندتعبیره‌ای MICE، می‌تواند دقت بازسازی داده‌های اقلیمی گمشده را به‌طور معناداری افزایش دهد. بر اساس شاخص‌های آماری R^2، NRMSE، |PBIAS| و KGE، مدل‌های ترکیبی نسبت به مدل پایه MICE عملکرد دقیق‌تر، پایدارتر و با بایاس کمتر از خود نشان دادند. در میان مدل‌های مورد بررسی، MICE–XGB بالاترین مقادیر R^2 (بیش از ۰/۹۰) و KGE (بیش از ۰/۹۲) را در اکثر ایستگاه‌ها کسب کرد و به‌عنوان کارآمدترین مدل در بازسازی داده‌ها شناخته شد. همچنین، مقادیر بسیار پایین |PBIAS| (کمتر از ۰/۰۲۵ درصد) نشان‌دهنده عدم وجود سوگیری سیستماتیک و صحت بازسازی‌ها بود.

از نظر متغیرهای اقلیمی، بهترین عملکرد مدل‌ها در بازسازی داده‌های دمایی و فشار بخار آب مشاهده شد، در حالی‌که بیشترین خطاها مربوط به متغیرهای وابسته به ابرناکی و تبخیر و تعریق بود. این تفاوت عملکرد به دلیل نوسانات زمانی بالا، پراکندگی داده و وابستگی شدید این متغیرها به شرایط لحظه‌ای و فصلی جوی است که بازسازی دقیق آن‌ها را دشوار می‌سازد. تحلیل مکانی نشان داد که ایستگاه‌های پایین‌ارتفاع و اقلیم معتدل‌تر (ارومیه، تبریز و بناب) دارای عملکرد بالاتری هستند، در حالی‌که ایستگاه‌های مرتفع و سرد (سقز و سراب) به دلیل نوسانات شدید روزانه و پراکندگی بالای داده‌ها، خطای بازسازی اندکی بیشتر دارند. این یافته‌ها تأکید می‌کند که توپوگرافی و ویژگی‌های اقلیمی هر ایستگاه نقش مهمی در عملکرد مدل‌ها ایفا می‌کنند.

برای ارزیابی عدم‌قطعیت و پایداری مدل‌ها، چارچوب چندتعبیره‌ای MICE به‌طور ذاتی امکان تولید چندین جایگزین برای داده‌های گمشده را فراهم می‌کند. این ویژگی ظرفیت روش پیشنهادی را برای بازنمایی دامنه عدم‌قطعیت مرتبط با فرآیند بازسازی داده‌ها نشان می‌دهد، هرچند در این پژوهش تمرکز اصلی بر مقایسه دقت میانگین بازسازی‌ها در سطح گمشدگی ۲۰٪ بوده است.

نتایج نشان داد که مدل‌های ترکیبی مبتنی بر الگوریتم‌های تقویتی، به‌ویژه MICE–XGB و MICE–LGBM، در این سطح گمشدگی عملکردی پایدار، دقیق و با بایاس بسیار کم در تمامی ایستگاه‌ها و متغیرهای اقلیمی از خود نشان داده‌اند. ثبات نتایج در ایستگاه‌های مختلف و شرایط مکانی–اقلیمی متفاوت، بیانگر استحکام و قابلیت تعمیم این مدل‌ها در بازسازی داده‌های اقلیمی است. همچنین، ارزیابی عملکرد مدل‌ها در مقیاس زمانی روزانه نشان‌دهنده پایداری آن‌ها در شرایط زمانی مختلف بوده است.

با وجود نتایج قابل‌قبول و عملکرد پایدار مدل‌های ترکیبی پیشنهادی، این پژوهش با برخی محدودیت‌ها همراه است. نخست آن‌که ارزیابی عملکرد مدل‌ها صرفاً در سطح گمشدگی ۲۰ درصد انجام شده است و رفتار مدل‌ها در سطوح پایین‌تر یا بالاتر گمشدگی مورد بررسی قرار نگرفته است. دوم، اگرچه چارچوب MICE امکان تولید چندین جایگزین و تحلیل عدم‌قطعیت را فراهم می‌کند، در این مطالعه تمرکز اصلی بر مقایسه دقت میانگین بازسازی‌ها بوده و بازه‌های اطمینان و عدم‌قطعیت کمی به‌صورت صریح محاسبه نشده‌اند. همچنین، تحلیل‌ها به داده‌های ایستگاهی روزانه محدود بوده و اثر مقیاس‌های زمانی بلندمدت‌تر و داده‌های مکمل نظیر داده‌های ماهواره‌ای و بازتحلیل در این پژوهش لحاظ نشده است. علاوه بر این، هزینه محاسباتی بالاتر برخی مدل‌های تقویتی می‌تواند کاربرد آن‌ها را در پایگاه‌های داده بسیار بزرگ با محدودیت‌های سخت‌افزاری با چالش مواجه سازد.

پیشنهادها برای مطالعات آینده:

بررسی تحلیل حساسیت مدل‌ها نسبت به سطوح مختلف گمشدگی داده (برای مثال ۱۰ تا ۳۰درصد) و ارزیابی اثر آن بر دقت و پایداری بازسازی.

محاسبه صریح بازه‌های اطمینان و عدم‌قطعیت کمی با استفاده از قواعد Rubin در چارچوب MICE.

استفاده از نسخه‌های بهینه‌شده الگوریتم‌های تقویتی مانند CatBoost یا AdaBoost در ترکیب با MICE برای بررسی اثر ساختار داده و نوع گمشدگی.

ارزیابی عملکرد مدل‌ها در دوره‌های زمانی طولانی‌تر (ماهانه و سالانه) برای تحلیل پایداری زمانی روش‌ها.

به‌کارگیری داده‌های ماهواره‌ای و بازتحلیل (مانند ERA5) در کنار داده‌های ایستگاهی برای افزایش دقت بازسازی.

استفاده از روش‌های یادگیری عمیق مانند LSTM یا Autoencoder در چارچوب چندتعبیره‌ای MICE برای ارتقای توان بازسازی.

ملاحظات اخلاقی

حامی مالی

نویسندگان هیچ بودجه خاصی برای این کار دریافت نکرده‌اند.

مشارکت در تألیف

مفهوم‌سازی، M.J. و M.Sh.؛ روش‌شناسی، M.J. و M.Sh.؛ نرم‌افزار، M.Sh.؛ اعتبارسنجی، .M.J، M.Sh. و S.E.؛ تحلیل رسمی، M.J.؛ تحقیق، M.J. و M.Sh.؛ منابع، M.J. و S.E.؛ گردآوری داده‌ها، M.J.؛ نگارش - تهیه پیش‌نویس اصلی، M.J.؛ نگارش - بررسی و ویرایش، M.Sh. و S.E. همه نویسندگان نسخه منتشر شده مقاله را خوانده و با آن موافقت کرده‌اند.

اعلامیه هوش مصنوعی مولد و فناوری‌های مبتنی بر هوش مصنوعی در فرآیند نگارش

نویسندگان اعلام می‌کنند که هیچ گونه هوش مصنوعی مولد یا فناوری‌های مبتنی بر هوش مصنوعی در نگارش، تحلیل یا تهیه این مقاله استفاده نشده است. نویسندگان مسئولیت کامل محتوای این نشریه را بر عهده دارند.

بیانیه دسترسی به داده‌ها

داده‌ها بنا به درخواست نویسندگان در دسترس هستند.

سپاسگزاری

از سازمان هواشناسی ایران به خاطر ارائه داده‌های تاریخی تشکر می‌کنیم. نویسندگان از داوران ناشناس به خاطر نظرات و پیشنهادات ارزشمندشان تشکر می‌کنند.

پیروی از اصول اخلاق پژوهش

نویسندگان اصول اخلاقی را در انجام و انتشار این پژوهش علمی رعایت نموده‌اند و این موضوع مورد تأیید همه آنهاست.

تعارض منافع

بنا بر اظهار نویسندگان این مقاله تعارض منافع ندارد.

REFERENCES

Afrifa-Yamoah, E., Mueller, U. A., Taylor, S. M., & Fisher, A. J. (2020). Missing data imputation of high-resolution temporal climate time series data. Meteorological Applications, 27(1), e1873. https://doi.org/https://doi.org/10.1002/met.1873

Alejo-Sanchez, L. E., Márquez-Grajales, A., Salas-Martínez, F., Franco-Arcega, A., López-Morales, V., Acevedo-Sandoval, O. A., González-Ramírez, C. A., & Villegas-Vega, R. (2025). Missing data imputation of climate time series: A review. MethodsX, 15, 103455. https//:doi.org/10.1016/j.mex.2025.103455

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? International Journal of Methods in Psychiatric Research, 20(1), 40-49. https//:doi.org/https://doi.org/10.1002/mpr.329

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA. https//:doi.org/10.1145/2939672.2939785

Costa, T., Falcão, B., Mohamed, M. A., Annuk, A., & Marinho, M. (2024). Employing machine learning for advanced gap imputation in solar power generation databases. Sci Rep, 14(1), 23801. https://doi.org/10.1038/s41598-02–74342-2

Davari, S., Eslamian, S., Jamali, M., & Safavi, H. R. (2025). Application of Machine Learning Algorithms for Groundwater Level Prediction in the Najafabad Plain. Sci Rep .

Farzandi, M., Sanaeinejad, H., Rezaei-Pazhan, H., & Sarmad, M. (2022). Improving estimation of missing data in historical monthly precipitation by evolutionary methods in the semi-arid area. Environment, Development and Sustainability, 24(6), 8313-8332. https://doi.org/10.1007/s10668-021-01784-4

Fazel Najafabadi, E., & Shayannejad, M. (2025). Evaluation the efficiency of machine learning boosting methods for estimating the water quality index of the Zayandeh Rood River. Iranian Journal of Soil and Water Research, 56(5), 1355-1378. https://doi.org/10.22059/ijswr.2025.392173.-669906

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics, 29(5), 1189-1232. http://www.jstor.org/stable/2699986

Golkhatmi, N. S. N., & Farzandi, M. (2024). Enhancing Rainfall Data Consistency and Completeness: A Spatiotemporal Quality Control Approach and Missing Data Reconstruction Using MICE on Large Precipitation Datasets. Water Resources Management, 38(3), 815-833. https://doi.org/10.1007/s11269-023-03567-0

Gupta Hoshin, V., Sorooshian, S & ،.Yapo Patrice, O. (1999). Status of Automatic Calibration for Hydrologic Models: Comparison with Multilevel Expert Calibration. Journal of Hydrologic Engineering, 4(2), 135-143. https://doi.org/10.1061/(ASCE)1084-0699(1999)4:2

Gupta, H. V., Kling ،H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. Journal of Hydrology, 377(1), 80-91. https://doi.org/https://doi.org/10.1016/j.jhydrol.20

Hasanpour Kashani, M., & Dinpashoh, Y. (2012). Evaluation of efficiency of different estimation methods for missing climatological data. Stochastic Environmental Research and Risk Assessment, 26(1), 59-71. https://doi.org/10.1007/s00477-011-053

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M . . . ،.Thépaut, J.-N. (2020). The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society, 146(730), 1999-2049. https://doi.org/https://doi.org/10.1002/qj.3803

Hosseinpour, S., Sharafati, A., & Abghari, H. (2025). Downscaling of two selected GCM data using a hybrid deep learning method of Wavelet-CNN-LSTM in Iran. Theoretical and Applied Climatology, 156(9): 459. DOI:10.1007/s00704-025-05685-8

Jääskeläinen, E., Manninen, T., Hakkarainen, J., & Tamminen, J. (2022). Filling gaps of black-sky surface albedo of the Arctic sea ice using gradient boosting and brightness temperature data. International Journal of Applied Earth Observation and Geoinformation, 107, 102701. https://doi.org/https://doi.org/10.1016/j.jag.2022.1

Davari, S., Elamian, S., Jamali, M., & Safavi, H. R. (2025). Application of machine learning algorithms for groundwater level prediction in the Najafabad plain. Scientific Reports, 14(3), 743-752. https://doi.org/10.1038/s41598-025-32376-1

Jamali Jezeh, M., Shayannejad, M., & Hejazi, S. M. (2020). Evaluation the Performance of Filters Made of BC, PET and

PP Textiles in Removing Oil Contaminants from Water [Research]. Journal of Water and Soil Science, 24(4), 295-312. https://doi.org/10.47176.jwss.24.4.42931

Jamali, M., Gohari, A., & Akhavan Saraf, G. (2024). Spatiotemporal evaluation of temperature and precipitation extremes indices over Iran under the influence of climate change. Water and Irrigation Management, 14(3), 743-752. https://doi.org/10.22059/jwim.2024.374814.1156

Jamali, M., Eslamian, S., Shayannejad, M., & Gohari, A. (2026). Observed warming–driven aridification and climate-type transitions across Iran. Journal of Arid Environments, 19(2), -. https://doi.org/10.1016/j.jaridenv.2026.105606

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30 .

Khosravi, G., Nafarzadegan, A. R., Nohegar, A., Fathizadeh, H., & Malekian, A. (2015). A modified distance-weighted approach for filling annual precipitation gaps: application to different climates of Iran. Theoretical and Applied Climatology, 119(1), 33-42. https://doi.org/10.1007/s00704-014-1091-5

Knoben, W. J. M., Freer, J. E., & Woods, R. A. (2019). Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. Hydrol. Earth Syst. Sci., 23(10), 4323-4331. https://doi.org/10.5194/hess-23-4323-2019

Legates, D. R., & McCabe Jr, G. J. (1999). Evaluating the use of "goodness-of-fit" Measures in hydrologic and hydroclimatic model validation. Water Resources Research, 35(1), 233-241. https://doi.org/https://doi.org/10.1029/1998WR900018

Little, R., & Rubin ،D. (1987). Multiple imputation for nonresponse in surveys. Wiley, 10, 9780470316696 .

Matinzadeh, M. m., Fattahi, R., Shayanzadeh, M., & Abdollahi, K. (2013). Estimation and Reconstruction of Annual Maximum 24-H Rainfall Data Using Combination of Genetic Algorithm and Artificial Neural Networks Models (Case Study: Chaharmahal va Bakhtiyari Province) .ijwmse, 7(22), 53. http://jwmsei.ir/article-1-245-fa.html

N. Moriasi, D., G. Arnold, J., W. Van Liew, M., L. Bingner, R., D. Harmel, R., & L. Veith, T. (2007). Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. Transactions of the ASABE, 50(3), 885-900. https://doi.org/https://doi.org/10.13031/2013.23153

Plein, M., Feigel, G., Zeeman, M., Dormann, C. F., & Christen, A. (2025). Using Gradient Boosting for gap-filling to analyze temperature and humidity patterns in an urban weather station network in Freiburg, Germany. Urban Climate, 62, 102496. https://doi.org/https://doi.org/10.1016/j.uclim.2025.102496

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. Psychological Methods, 7(2), 147-177. https://doi.org/10.1037/1082-989X.7.2.147

Van Buuren, S. (2000). Multivariate imputation by chained equations: MICE V1. 0 user's manual. Leiden: TNO .

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. Journal of statistical software, 45, 1-67 .

Willmott, C. J. (1981). ON THE VALIDATION OF MODELS. Physical Geography, 2(2), 184-194. https://doi.org/10.1080/02723646.1981.10642213