

Evaluation of MICE-Based Machine Learning Models for Reconstructing Missing Climate Data in the Urmia Lake Basin

Mohammad Shayannejad^{1*} , Mohammad Jamali¹ 

¹ Department of Water Science and Engineering, College of Agriculture, Isfahan University of Technology, Isfahan, Iran

Extended Abstract

Introduction

observational climate records frequently suffer from missing values due to instrument malfunction, station relocation, data transmission errors, or long-term interruptions in measurements. If not appropriately addressed, missing data can introduce bias, reduce statistical power, and compromise the reliability of subsequent modeling and decision-making processes. This challenge is particularly critical in regions with complex climatic variability and environmental sensitivity, such as the Lake Urmia Basin in northwestern Iran. Traditional approaches for handling missing climatic data, including listwise deletion or simple statistical substitution (e.g., mean or median imputation), are computationally convenient but often distort the statistical structure of the data and fail to capture inter-variable dependencies. In response to these limitations, advanced multivariate and machine-learning-based imputation methods have gained increasing attention. Among them, Multiple Imputation by Chained Equations (MICE) has emerged as a robust framework that accounts for uncertainty and exploits relationships among multiple variables. Recent studies suggest that integrating MICE with machine learning algorithms can further enhance imputation accuracy, particularly for non-linear and highly interdependent climatic variables. Nevertheless, comprehensive evaluations comparing different MICE-based hybrid models across multiple climatic variables and stations remain limited. Therefore, this study aims to systematically assess and compare the performance of standard MICE and four hybrid approaches MICE-Linear Regression (MICE-LR), MICE- Decision Tree (MICE-DT), MICE-K-Nearest Neighbor (MICE-KNN), and MICE- Support Vector Machine (MICE-SVM), across a wide range of climatic variables and meteorological stations within the Lake Urmia Basin.

Materials and Methods

This study was conducted using daily climatic data from six synoptic meteorological stations located in the Lake Urmia Basin. The dataset includes a diverse set of climatic variables representing thermal conditions, atmospheric moisture, cloudiness, wind characteristics, radiation and energy balance, and sea-level pressure. To ensure consistency and robustness, all variables were preprocessed through quality control procedures, including outlier detection and temporal consistency checks. Missing data were reconstructed using five imputation approaches: standard MICE and four hybrid MICE-based models (MICE-LR, MICE-DT, MICE-KNN, and MICE-SVM). The imputation procedure was implemented iteratively within the chained equations framework to ensure convergence and stability of the reconstructed values. Model performance was evaluated using multiple complementary statistical metrics, including the coefficient of determination (R^2), normalized root means square error (NRMSE), Kling-Gupta Efficiency (KGE), and percent bias (PBIAS).

Results and Discussion

The comparative analysis revealed substantial differences in imputation performance among the evaluated models, depending on the type of climatic variable and station characteristics. Overall, hybrid MICE-based models demonstrated superior performance compared to the standard MICE approach, particularly for temperature-related variables and atmospheric moisture parameters. Among the hybrid models, MICE-DT achieved comparatively higher KGE values for several variables, highlighting its ability to model non-linear interactions. Nevertheless, both MICE-DT and MICE-LR provided a more balanced trade-off between reconstruction accuracy and computational efficiency. In contrast, MICE-KNN and MICE-SVM exhibited variable performance, with notable sensitivity to station-specific conditions and variable type. While these models performed reasonably well for certain variables, their performance deteriorated for others, especially in cases involving higher variability or weaker spatial coherence. Standard MICE and MICE-LR showed comparable results, suggesting that linear assumptions may be insufficient for fully representing the dynamics of complex climatic systems. The station-

based analysis highlighted spatial heterogeneity in model performance, emphasizing the influence of local climatic and topographic conditions. Furthermore, the computational analysis indicated that while hybrid models generally required longer execution times than standard MICE, MICE–DT provided a favorable balance between accuracy and computational efficiency.

Conclusion

This study provides a comprehensive evaluation of standard and hybrid MICE-based imputation methods for reconstructing missing climatic data in a multi-variable and multi-station framework. The results demonstrate that incorporating machine learning algorithms within the MICE framework substantially improves reconstruction accuracy, particularly for variables characterized by non-linear behavior. Among the evaluated models, MICE–DT emerged as the most robust and efficient approach, offering consistently high performance across different climatic variables and stations. Despite these strengths, certain limitations were identified. The performance of some hybrid models, particularly MICE–KNN and MICE–SVM, showed sensitivity to station-specific conditions and increased computational demand, which may limit their applicability in large-scale studies. These findings suggest that no single imputation method is universally optimal, and model selection should be tailored to the characteristics of the dataset and research objectives. From a practical perspective, the proposed framework provides valuable guidance for researchers and practitioners seeking reliable methods for handling missing climatic data. The results have direct implications for hydrological modeling, climate trend analysis, and environmental impact assessments in data-scarce regions.

Keywords: Cross-validation, Data uncertainty, Decision tree model, Incomplete meteorological data, Lake Urmia Basin, MICE

Article Type: Research Article

Acknowledgement

We would like to express our sincere gratitude to Isfahan University of Technology for their financial and logistical support, which significantly contributed to the success of this research project.

Conflicts of Interest

The authors of this article declare no conflicts of interest regarding the authorship or publication of this article.

Data Availability Statement

The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Authors' contribution

Mohammad Shayannejad: Supervision, Methodology validation, Results interpretation, Review and editing.

Mohammad Jamali: Conceptualization, Data processing, Modeling, Software analysis, Writing – original draft.

*Corresponding Author, E-mail: m_yazdani@semnan.ac.ir

Citation: Shayannejad, M., & Jamali, M. (2026). Evaluation of MICE-Based Machine Learning Models for Reconstructing Missing Climate Data in the Urmia Lake Basin. *Water and Soil Management and Modelling*, 6(3), 181-199

doi: 10.22098/mmws.2026.19270.1777

Received: 26 January 2026, Received in revised form: 10 March 2026, Accepted: 09 April 2026, Published online: 27 August 2026.

Water and Soil Management and Modeling, Year 2026, Vol. 6, No. 3, pp. 181-199.

Publisher: University of Mohaghegh Ardabili

© Author(s)





ارزیابی مدل های یادگیری ماشین مبتنی بر MICE در بازسازی داده های گمشده اقلیمی در حوضه دریاچه ارومیه

محمد شایان نژاد^{۱*}، محمد جمالی^۱

^۱ گروه علوم و مهندسی آب، دانشکده کشاورزی، دانشگاه صنعتی اصفهان، اصفهان، ایران

چکیده

وجود داده های گمشده در شبکه های ایستگاه های هواشناسی یکی از چالش های مهم در مطالعات هیدرولوژیکی و اقلیمی است که می تواند موجب افزایش عدم قطعیت و کاهش دقت تحلیل ها شود. انتخاب روش مناسب برای بازسازی این داده ها نقش تعیین کننده ای در بهبود کیفیت نتایج دارد. در این پژوهش، عملکرد روش MICE و چهار رویکرد ترکیبی آن شامل MICE-LR، MICE-DT، MICE-KNN و MICE-SVM در بازسازی داده های گمشده متغیرهای اقلیمی مورد ارزیابی قرار گرفت. تحلیل ها با استفاده از داده های شش ایستگاه منتخب حوضه دریاچه ارومیه انجام شد که تنوع مکانی و اقلیمی مناسبی را پوشش می دهند. ارزیابی عملکرد مدل ها بر اساس شاخص کارایی کلینگ-گوپتا (KGE)، درصد اریبی (PBIAS) و زمان اجرای محاسباتی صورت گرفت. نتایج نشان داد که تمامی مدل ها توانایی قابل قبولی در بازسازی داده های گمشده دارند؛ با این حال، بهبود عملکرد نسبت به روش پایه MICE وابسته به نوع مدل بوده است. در این میان، مدل های MICE-DT و MICE-SVM عملکرد برتری نسبت به مدل پایه از خود نشان دادند؛ به طوری که مدل MICE-DT در بازسازی متغیرهای دمای، ابرناکی و فشار و مدل MICE-SVM در متغیرهای رطوبتی دقت بالاتری را ارائه کردند. با این وجود، روش پایه MICE در متغیرهای تابش و انرژی عملکردی قابل رقابت با مدل های ترکیبی داشت. در این میان، مدل های MICE-LR و MICE-DT با مقادیر کمتر PBIAS و زمان اجرای پایین تر، توازن مناسبی بین دقت بازسازی و کارایی محاسباتی برقرار کردند. در مقابل، مدل MICE-KNN به دلیل نیاز به محاسبه فاصله بین نمونه ها در هر تکرار، زمان اجرای بیشتری نسبت به مدل های خطی داشت. همچنین، مدل MICE-SVM به علت فرآیند بهینه سازی تکرارشونده در آموزش و تنظیم پارامترهای کرنل در هر مرحله از الگوریتم MICE، زمان اجرای قابل توجهی را به خود اختصاص داد؛ از این رو، در داده های حجیم و کاربردهای عملی در مقیاس های بزرگ گزینه ای کم کارآمدتر محسوب می شود. به طور کلی، نتایج این مطالعه نشان می دهد که استفاده از رویکردهای ترکیبی ساده تر مبتنی بر MICE می تواند راهکاری مؤثر و کارآمد برای بازسازی داده های گمشده اقلیمی باشد.

واژه های کلیدی: اعتبارسنجی متقابل، حوضه دریاچه ارومیه، داده های ناقص هواشناسی، عدم قطعیت داده ها، مدل درخت تصمیم، MICE

نوع مقاله: پژوهشی

*مسئول مکاتبات، پست الکترونیکی: shayannejad@iut.ac.ir

استناد: شایان نژاد، محمد و جمالی، محمد (۱۴۰۵). ارزیابی مدل های یادگیری ماشین مبتنی بر MICE در بازسازی داده های گمشده اقلیمی در حوضه دریاچه ارومیه. مدل سازی و مدیریت آب و خاک، ۶(۳)، ۱۸۱-۱۹۹.

doi:10.22098/mmws.2026.19270.1777

تاریخ دریافت: ۱۴۰۴/۱۱/۰۶ تاریخ بازنگری: ۱۴۰۴/۱۲/۱۹ تاریخ پذیرش: ۱۴۰۵/۰۱/۲۰ تاریخ انتشار: ۱۴۰۵/۰۶/۰۵

مدل سازی و مدیریت آب و خاک، سال ۱۴۰۵، دوره ۶، شماره ۳، صفحه ۱۸۱ تا ۱۹۹.

© نویسندگان

ناشر: دانشگاه محقق اردبیلی



۱- مقدمه

داده‌های اقلیمی پایه و اساس بسیاری از مطالعات مرتبط با هیدرولوژی، تغییر اقلیم، مدیریت منابع آب، کشاورزی و ارزیابی مخاطرات طبیعی را تشکیل می‌دهند. کیفیت، پیوستگی و کامل بودن این داده‌ها نقش تعیین‌کننده‌ای در اعتبار تحلیل‌ها، مدل‌سازی‌ها و تصمیم‌گیری‌های مدیریتی دارد (Liu et al., 2021; Jamali & Eslamian, 2024). با این حال، شبکه‌های ایستگاه‌های هواشناسی در بسیاری از مناطق جهان، به‌ویژه در کشورهای در حال توسعه، به‌طور گسترده با مسئله داده‌های گمشده مواجه‌اند (Diouf et al., 2023; Jamali & Eslamian, 2023). این کمبود داده می‌تواند ناشی از نقص تجهیزات اندازه‌گیری، خطاهای انسانی، جابه‌جایی ایستگاه‌ها، وقفه‌های طولانی در ثبت داده‌ها یا شرایط نامساعد محیطی باشد. وجود داده‌های گمشده، در صورت عدم مدیریت صحیح، می‌تواند منجر به سوگیری آماری، کاهش دقت مدل‌ها و تضعیف قابلیت اعتماد نتایج شود (Little & Rubin, 2019; O'Sullivan & Kelly, 2024, Vafaei & Mamaghani et al., 2026). در دهه‌های گذشته، روش‌های متعددی برای برخورد با داده‌های گمشده پیشنهاد شده است. ساده‌ترین رویکردها شامل حذف رکوردهای ناقص یا جایگزینی داده‌های گمشده با مقادیر میانگین یا میانه هستند. اگرچه این روش‌ها از نظر محاسباتی ساده‌اند، اما اغلب منجر به از دست رفتن اطلاعات ارزشمند، کاهش واریانس داده‌ها و ایجاد سوگیری در نتایج می‌شوند. روش‌های آماری پیشرفته‌تری مانند رگرسیون خطی، رگرسیون چندگانه و تحلیل مؤلفه‌های اصلی (PCA^1) نیز مورد استفاده قرار گرفته‌اند، اما این روش‌ها عمدتاً بر فرضیات خطی و ساختارهای ساده همبستگی متکی‌اند و در مواجهه با روابط پیچیده و غیرخطی موجود در داده‌های اقلیمی با محدودیت مواجه می‌شوند (Matinzadeh et al., 2013; Alejo-Sanchez et al., 2025). Matinzadeh et al. (2013) به تخمین و بازسازی داده‌های حداکثر بارش ۲۴ ساعته در استان چهارمحال و بختیاری با استفاده از رگرسیون فازی پرداختند. برای ارزیابی مدل از شاخص‌های آماری RMSE و R^2 استفاده کردند و نتایج حاصل از مدل رگرسیون فازی با روش‌های نسبت نرمال، محورهای مختصات، رگرسیون ساده و رگرسیون چند متغییره مقایسه شد. در این پژوهش نتایج بازسازی در هر چهار اقلیم مختلف حاکی از برتری معنی‌دار روش رگرسیون فازی نسبت به سایر روش‌ها بود. Sahoo and Ghose (2022) در مطالعه‌ای به بازسازی داده‌های مفقودی در ۱۹ ایستگاه در حوضه کاجار در کشور هند پرداختند. در

این تحقیق از روش‌های KNN، نقشه‌های خودسازمانده^۲، جنگل تصادفی^۳ و شبکه عصبی پیشخور^۴ استفاده شد و شاخص‌های RMSE، R^2 و MAE برای مقایسه روش‌ها محاسبه شد. برای روش شبکه عصبی پیشخور مقادیر MAE، R^2 و RMSE به ترتیب برابر ۰/۰۴۳، ۰/۹۹۹ و ۰/۰۶۶ بود که بهترین عملکرد را به‌ویژه در مناطق با کمبود شدید داده نشان داد. (Karbasi 2016) به بازسازی داده‌های مفقودی مجموع ماهیانه ساعت آفتابی مورد نیاز برای محاسبه تبخیر و تفرق در ایستگاه سینوپتیک زنجان با استفاده از دو نوع شبکه عصبی مصنوعی MLP و RBF پرداخت. نتایج تحقیق نشان داد که می‌توان با استفاده از داده‌های هواشناسی ایستگاه‌های هدف و ایستگاه‌های مجاور اقدام به بازسازی داده‌های مجموع ماهانه ساعت آفتابی با دقت بالا کرد. مقایسه عملکرد بین دو شبکه عصبی نشان داد که دقت شبکه عصبی MLP تا حدودی بیشتر از دیگری است. (Mohtasham 2023) به بازسازی داده‌های بارندگی روزانه دارای دوره‌های مفقودی با استفاده از روش‌های محاسبه چندگانه با تطبیق میانگین پیش‌بینی و جنگل تصادفی برای ۳۰ ایستگاه باران‌سنجی موجود در دشت مشهد-چناران پرداخت. در این پژوهش تحلیل حوضه‌ای توسط روش سلسله مرتبی Ward انجام شد و نتایج نشان داد که روش جنگل تصادفی در اکثر خوشه‌ها از عملکرد بهتری نسبت به روش تطبیق میانگین پیش‌بینی برخوردار است. (Plein et al 2025) با استفاده از مدل $XGBoost^5$ داده‌های ناقص دما و رطوبت ایستگاه‌های هواشناسی فرایبورگ آلمان را بازسازی کردند. این مدل با شکاف‌های داده‌ای ۱ تا ۲۸ روزه، دقت بالایی نشان داد (RMSE دما: ۰/۴۶ کلوین و رطوبت: ۲/۵۱ درصد). حساسیت مدل به تعداد داده‌های ناقص کم بود، اما در ایستگاه‌های دور از شهر دقت کاهش یافت. نتایج بازسازی نشان داد مرکز شهر ۱/۱ درجه گرم‌تر و دارای ۷ درصد فشار بخار کمتر نسبت به مناطق روستایی است. این تحقیق کارایی بالای $XGBoost$ را در تکمیل داده‌های هواشناسی شهری تایید کرد. (Jääskeläinen et al 2022) نیز از گرادیان بوستینگ برای بازسازی داده‌های آلبیدوی یخ‌های دریایی قطب شمال (به دلیل پوشش ابری و زاویه تابش) استفاده کردند. این مدل با میانگین ماهانه آلبیدو، دما و غلظت یخ آموزش دید و با RMSE حدود ۰/۰۴۸ و خطای نسبی کمتر از ۲۰ درصد، عملکرد دقیق‌تر و پایدارتری نسبت به رگرسیون خطی به‌ویژه در پیش‌بینی آلبیدوی یخ‌های در حال ذوب داشت.

⁴ Feed-Forward Neural Network

⁵ Extreme Gradient Boosting

¹ Principal Component Analysis

² Self-Organizing Maps

³ Random Forest

تشخیص داده‌های پرت به کار گرفته شد. مرور مطالعات پیشین نشان می‌دهد که با وجود بررسی‌های گسترده روش‌های بازسازی داده‌های مفقود، همچنان کمبودهای قابل توجهی در این زمینه وجود دارد. بسیاری از مطالعات تنها بر یک یا دو متغیر اقلیمی تمرکز کرده‌اند، یا ارزیابی‌ها را به یک ایستگاه یا یک شاخص عملکرد محدود نموده‌اند (Jing et al., 2022; Diouf et al., 2023; Golkhatmi & Farzandi, 2024; Alejo-Sanchez, 2025; Kumar, G.P., 2025). هم‌چنین، مقایسه هم‌زمان چند مدل ترکیبی MICE در مقیاس چندایستگاهی و برای مجموعه‌ای متنوع از متغیرهای اقلیمی، به‌ویژه در مناطق حساس اقلیمی مانند حوضه دریاچه ارومیه، کمتر مورد توجه قرار گرفته است. با توجه به مطالب فوق، هدف اصلی این پژوهش ارزیابی جامع و مقایسه عملکرد روش استاندارد MICE و چهار رویکرد ترکیبی شامل MICE-SVM، MICE-KNN، MICE-DT، MICE-LR در بازسازی داده‌های گمشده اقلیمی است. این ارزیابی برای مجموعه‌ای متنوع از متغیرهای اقلیمی شامل دما، ابرناکی، باد، رطوبت و بخار آب، تابش و انرژی و فشار سطح دریا و در شش ایستگاه منتخب حوضه دریاچه ارومیه انجام شده است. در این مطالعه، عملکرد مدل‌ها با استفاده از شاخص‌های آماری مکمل شامل R^2 ، NRMSE، KGE و PBIAS مورد بررسی قرار گرفته و علاوه بر دقت بازسازی، کارایی محاسباتی مدل‌ها از طریق تحلیل زمان اجرا نیز ارزیابی شده است. نوآوری اصلی این پژوهش در ارائه یک چارچوب مقایسه‌ای چندبعدی (متغیرمحور، مدل‌محور و ایستگاه‌محور) نهفته است که می‌تواند به عنوان راهنمایی عملی برای انتخاب مناسب‌ترین روش بازسازی داده‌های گمشده اقلیمی در مطالعات مشابه مورد استفاده قرار گیرد.

۲- مواد و روش‌ها

۲-۱- منطقه مطالعاتی

حوضه آبریز دریاچه ارومیه به عنوان بزرگ‌ترین حوضه بسته ایران و یکی از مهم‌ترین زیست‌بوم‌های نیمه‌خشک منطقه خاورمیانه، در شمال غرب کشور و در محدوده سه استان آذربایجان شرقی، آذربایجان غربی و کردستان واقع شده است. مساحت این حوضه حدود ۵۱ هزار کیلومتر مربع برآورد می‌شود و به دلیل تنوع توپوگرافی، گستره ارتفاعی وسیع و تغییرات مکانی بارش و دما، از پیچیدگی اقلیمی قابل توجهی برخوردار است (Hassanzadeh et al., 2012). این ویژگی‌ها، همراه با نقش اقتصادی و اجتماعی منطقه، اهمیت داده‌های اقلیمی دقیق و پایدار را برای مدیریت منابع آب، تحلیل خشکسالی، و پایش تغییرات محیطی دوچندان

در این میان، روش $MICE^1$ به عنوان یکی از پرکاربردترین و مؤثرترین رویکردهای چندمتغیره برای بازسازی داده‌های گمشده معرفی شده است. مزیت اصلی MICE در آن است که با استفاده از روابط متقابل میان متغیرها، فرآیند بازسازی را به صورت تکرارشونده و مبتنی بر عدم قطعیت انجام می‌دهد و در نتیجه، ساختار آماری داده‌ها را بهتر حفظ می‌کند (Van Buuren, 2000; Azur et al., 2011). مطالعات نشان داده‌اند که روش MICE نسبت به روش‌های تک‌جایگزینی، دقت و پایداری بالاتری در بازسازی داده‌های اقلیمی دارد، به‌ویژه زمانی که درصد داده‌های گمشده قابل توجه باشد (Hamzah et al., 2022; Kumar & Dwarakish, 2025). با پیشرفت روش‌های یادگیری ماشین، استفاده از الگوریتم‌های داده‌محور برای بازسازی داده‌های گمشده به‌طور چشمگیری افزایش یافته است. الگوریتم‌هایی مانند KNN^2 ، SVM^4 و DT^3 به دلیل توانایی در مدل‌سازی روابط غیرخطی و تعاملات پیچیده میان متغیرها، در مطالعات متعددی مورد توجه قرار گرفته‌اند (Chivers et al., 2020; Sattari et al., 2024; Ghanavati et al., 2025). نتایج مطالعات نشان می‌دهد که عملکرد این الگوریتم‌ها به شدت به نوع متغیر اقلیمی، ویژگی‌های مکانی ایستگاه‌ها و ساختار داده‌ها وابسته است؛ به‌طوری که هیچ روش واحدی را نمی‌توان به عنوان گزینه برتر مطلق برای تمامی شرایط معرفی کرد (Amekudzi et al., 2022). در سال‌های اخیر، رویکردهای ترکیبی که MICE را با الگوریتم‌های یادگیری ماشین تلفیق می‌کنند، به عنوان راهکاری نوین برای بهبود دقت بازسازی داده‌ها مطرح شده‌اند (Jing et al., 2022). در این چارچوب، مدل‌های مختلف یادگیری ماشین به عنوان هسته پیش‌بینی در مراحل زنجیره‌ای MICE به کار گرفته می‌شوند. مطالعات نشان داده‌اند که ترکیب MICE با الگوریتم‌هایی مانند DT ، SVM یا رگرسیون خطی می‌تواند منجر به کاهش خطا، افزایش ضریب همبستگی و بهبود شاخص‌های کارایی نسبت به MICE استاندارد شود (Hamzah et al., 2022; Jing et al., 2022). با این حال، عملکرد این مدل‌های ترکیبی همچنان به نوع متغیر، شرایط اقلیمی منطقه و ویژگی‌های ایستگاه‌ها وابسته است و نیاز به ارزیابی‌های جامع و منطقه‌محور دارد (Qaraghuli et al., 2024). Golkhatmi and Farzandi (2024) با هدف کنترل کیفی و بازسازی داده‌های ناقص بارش در حوضه قره‌قوم شمال شرقی ایران، از داده‌های ۱۴۱ ایستگاه استفاده کردند. در این تحقیق، ابتدا خطاهای عمده، ناسازگاری‌های زمانی و داده‌های پرت مورد شناسایی قرار گرفت. سپس متغیرهایی مانند تعداد روزهای بارش در هر ماه، بیشینه بارش ماهانه و بارش استانداردشده برای

³ Decision Tree

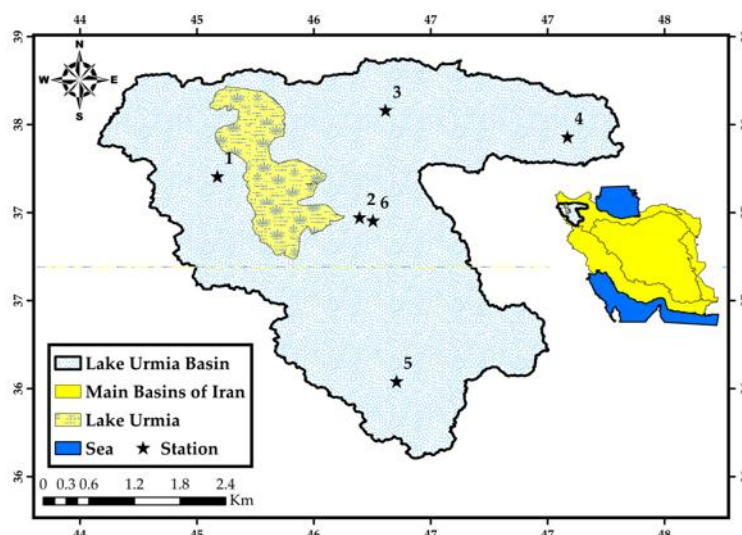
⁴ Support Vector Machine

¹ Multiple Imputation by Chained Equations

² K-Nearest Neighbors

مدل های مدیریتی و پیش بینی کننده دارد؛ از این رو بازسازی مناسب داده های گمشده یکی از چالش های مهم مطالعات مرتبط با این حوضه است. برای اجرای این مطالعه، شش ایستگاه سینوپتیک شاخص با پراکندگی جغرافیایی مناسب، شامل تبریز، بناب، ارومیه، مراغه، سقز و سراب، انتخاب شدند (جدول ۱ و شکل های ۱ و ۲).

می کند. در دهه های اخیر، افت شدید تراز آب دریاچه، کاهش رواناب های ورودی و بروز دوره های خشکسالی پی درپی، ضرورت بازسازی و تکمیل داده های اقلیمی منطقه را برای مطالعات هیدرولوژیکی و اقلیم شناسی آشکار ساخته است. دقت داده های بارش، دما و سایر عناصر اقلیمی تأثیر مستقیمی بر عملکرد



شکل ۱- موقعیت ایستگاه های سینوپتیک منتخب

Figure 1- Location of selected synoptic stations

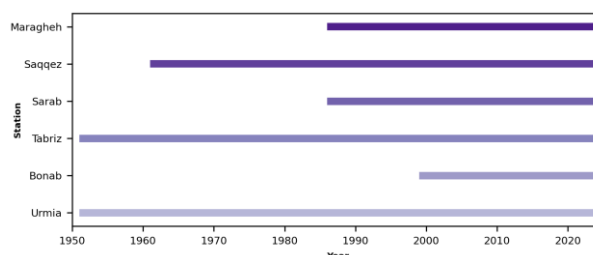
ایستگاه های منتخب در این پژوهش آورده شده است. شکل ۲ بیانگر سال آغاز بهره برداری، سال پایان و طول دوره آماری ایستگاه های منتخب است. ایستگاه های ارومیه و تبریز با دوره آماری ۱۹۵۱ تا ۲۰۲۵، طولانی ترین سری زمانی را در اختیار دارند که این موضوع می تواند در افزایش پایداری تحلیل های آماری و کاهش عدم قطعیت نتایج مؤثر باشد.

این ایستگاه ها نماینده طیفی از شرایط اقلیمی حوضه، از نواحی پست و نسبتاً گرم جنوب و جنوب غرب تا مناطق مرتفع و سرد شمال شرق و غرب هستند. اختلاف ارتفاع، تفاوت در الگوهای بارش و رژیم دمایی، و پراکندگی جغرافیایی این ایستگاه ها امکان ارزیابی عملکرد روش های بازسازی داده در شرایط متنوع محیطی و اقلیمی را فراهم می سازد. در جدول ۱ مشخصات جغرافیایی

جدول ۱- مشخصات جغرافیایی ایستگاه های منتخب

Table 1- Geographical characteristics of the selected stations

Num.	Station Name	Latitude (°E)	Longitude (°N)	Elevation (m)
1	Urmia	45.056	37.658	1315.9
2	Bonab	46.052	37.370	1290
3	Tabriz	46.224	38.122	1361
4	Sarak	47.508	37.935	1682
5	Saqqez	46.311	36.221	1522.8
6	Maragheh	46.146	37.348	1477.7



شکل ۲- مقایسه طول دوره آماری ایستگاه های سینوپتیک حوضه دریاچه ارومیه

Figure 2- Comparison of the statistical record lengths of synoptic stations in the Urmia Lake Basin

قرار گرفتند تا کیفیت، یکپارچگی و همسانی سری‌های زمانی تضمین شود. در ابتدا، داده‌ها از نظر مقادیر گمشده، مقادیر پرت و ناسازگاری‌های زمانی بررسی و اصلاح شدند و مقادیر گمشده ثبت گردید تا میزان واقعی داده‌های ناقص برای هر ایستگاه و متغیر محاسبه شود. داده‌های پرت با استفاده از معیارهای آماری و نمودارهای جعبه‌ای شناسایی و در صورت نیاز تصحیح یا حذف شدند. تمامی سری‌های زمانی به فرمت روزانه استانداردسازی شدند و داده‌های ناقص یا نامناسب اصلاح و یکنواخت شدند تا هر متغیر آماده اعمال روش‌های بازسازی گردد. برای ارزیابی عملکرد روش MICE و مدل‌های یادگیری ماشین، ۲۰ درصد از داده‌ها به‌طور تصادفی و عمدی حذف شدند تا شرایط گمشدگی شبیه‌سازی شود؛ این سطح همسو با محدوده واقعی گمشدگی داده‌ها در شبکه‌های اقلیم‌سنجی ایران و مطالعات پیشین انتخاب شد (Azur et al., 2022; Farzandi et al., 2011). داده‌ها پس از ایجاد گمشدگی مصنوعی به دو مجموعه بازسازی، شامل داده‌های ناقص برای اعمال مدل‌ها و ارزیابی شامل داده‌های واقعی حذف‌شده به‌عنوان مرجع سنجش دقت بازسازی تقسیم شدند.

این تنوع مکانی و اقلیمی، بستر مناسبی برای آزمون و مقایسه روش‌های مبتنی بر MICE-ML ایجاد می‌کند و می‌تواند قابلیت تعمیم‌پذیری نتایج را برای کاربردهای هیدرولوژیکی و اقلیم‌شناسی در مناطق نیمه‌خشک مشابه افزایش دهد؛ موضوعی که در مطالعات مرتبط با بازسازی داده‌های اقلیمی نیز مورد تأکید قرار گرفته است (Diouf et al., 2023; Kumar and Dwarakish, 2025).

۲-۲- جمع‌آوری و پیش‌پردازش داده‌ها

در این پژوهش، داده‌های اقلیمی روزانه مربوط به ۱۶ متغیر هواشناسی به‌عنوان ورودی اصلی مورد استفاده قرار گرفتند (جدول ۲). این مجموعه شامل پارامترهای دما، رطوبت نسبی، فشار، ساعت آفتابی، تبخیر، ابرناکی و مولفه‌های باد است که به‌صورت جامع نمایانگر شرایط اقلیمی متنوع حاکم بر حوضه آبریز دریاچه ارومیه می‌باشند. بررسی اولیه داده‌های خام نشان داد که تمامی متغیرها در دوره‌های زمانی مختلف دارای مقادیر گمشده هستند؛ بنابراین فرآیند بازسازی به‌صورت مستقل برای هر متغیر و با بهره‌گیری از روش MICE و نسخه‌های ترکیبی آن با مدل‌های یادگیری ماشین انجام گردید. قبل از بازسازی داده‌های گمشده، داده‌های اقلیمی جمع‌آوری‌شده از ایستگاه‌های منتخب تحت مراحل پیش‌پردازش

جدول ۲. متغیرهای اقلیمی روزانه مورد استفاده در این پژوهش

Table 2- Daily climatic variables used in this study

Category	Variable Name	Symbol	Unit	Category	Variable Name	Symbol	Unit
Humidity & Vapor	Maximum Relative Humidity	RHmax	%	Temperature	Maximum Air Temperature	Tmax	°C
	Minimum Relative Humidity	RHmin	%		Minimum Air Temperature	Tmin	°C
	Mean Relative Humidity	RHmean	%		Mean Air Temperature	Tmean	°C
	Mean Vapor Pressure	VPmean	hPa		Mean Dew Point Temperature	TDmean	°C
	Mean Saturation Vapor Pressure	SVPmean	hPa	Cloud Cover	Mean Cloud Cover	CLDmean	oktas
Radiation & Energy	Sunshine Duration	Sun	h		Maximum Cloud Cover	CLDmax	oktas
	Evapotranspiration	ET	mm	Wind	Maximum Wind Speed	WSmax	m/s
Pressure	Mean Sea Level Pressure	SLPmean	hPa		Minimum Wind Speed	WSmin	m/s

و KNeighborsRegressor، DecisionTreeRegressor و SVR استفاده شد.

۲-۳- روش‌های بازسازی داده‌ها

در این مطالعه، فرآیند بازسازی داده‌های گمشده اقلیمی از طریق دو رویکرد اصلی شامل روش آماری بازسازی چندگانه با معادلات زنجیره‌ای (MICE) و نیز ترکیب چارچوب MICE با الگوریتم‌های یادگیری ماشین انجام شد (جدول ۳). انتخاب چهار الگوریتم یادگیری ماشین در این پژوهش با هدف پوشش طیف متنوعی از

متغیرها در قالب ماتریسی سازماندهی شدند که مقادیر روزانه هر متغیر در هر ایستگاه را شامل می‌شد و امکان اجرای یکپارچه الگوریتم‌های MICE و مدل‌های یادگیری ماشین و مقایسه عملکرد بازسازی در ایستگاه‌ها و متغیرهای مختلف را فراهم ساخت. تمامی تحلیل‌ها در محیط Python (نسخه ۱۰/۸) انجام شد. برای پیاده‌سازی چارچوب MICE از کتابخانه IterativeImputer در بسته scikit-learn و برای مدل‌های یادگیری ماشین از ماژول‌های LinearRegression،

شرایط متفاوت ساختار داده و روابط بین متغیره را فراهم می کند. همچنین، این الگوریتم ها قابلیت پیاده سازی در چارچوب تکرار شونده MICE را داشته و از نظر پیچیدگی محاسباتی سطوح متفاوتی را نمایندگی می کنند که تحلیل توازن بین دقت و هزینه محاسباتی را ممکن می سازد.

ساختارهای مدل سازی انجام شد. به طور مشخص، LR به عنوان نماینده مدل های خطی، DT به عنوان الگوریتم غیر خطی مبتنی بر قواعد و آستانه ها، KNN به عنوان روش مبتنی بر فاصله و شباهت محلی، و SVM به عنوان مدل کرنلی با توانایی مدل سازی روابط پیچیده انتخاب شدند. این تنوع امکان مقایسه عملکرد مدل ها در

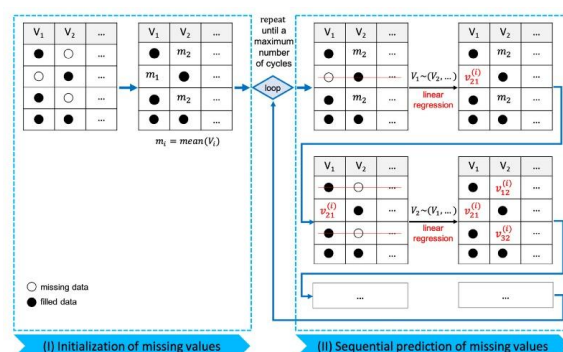
جدول ۳- مدل های آماری و ترکیبی مورد استفاده برای بازسازی داده های گمشده اقلیمی

Table 3- Statistical and hybrid models used for missing climate data imputation

Category	Model Name	Symbol	Description
Statistical	Multiple Imputation by Chained Equations	MICE	A multiple imputation method based on chained equations, enabling multivariate data reconstruction while preserving the correlation structure among variables.
	MICE + Linear Regression	MICE-LR	Integration of MICE with linear regression to predict missing values of each variable based on other observed variables.
Hybrid	MICE + Decision Tree Regression	MICE-DT	Application of decision tree regression as the prediction algorithm within the MICE framework for missing data reconstruction.
	MICE + K-Nearest Neighbors Regression	MICE-KNN	Reconstruction of missing values using the KNN algorithm within the MICE framework, leveraging the nearest neighboring samples.
	MICE + Support Vector Machine Regression	MICE-SVM	Use of Support Vector Machine (SVM) regression within the MICE framework to predict missing data and model nonlinear relationships.

معنا که برای هر داده گمشده چند مقدار جایگزین بر اساس توزیع شرطی سایر متغیرها تولید می شود. فرآیند بازسازی در قالب زنجیره ای از معادلات رگرسیونی تکرار شونده اجرا می شود؛ در هر گام، مقادیر گمشده یک متغیر با توجه به سایر متغیرها تخمین زده می شوند (شکل ۳).

روش MICE: این روش یکی از پرکاربردترین و معتبرترین روش ها برای بازسازی داده های گمشده در مطالعات آماری و علوم محیطی است (Little & Rubin, 1987; Van Buuren, 2000). این روش بر پایه رویکرد «چندجایگزینی» عمل می کند؛ به این



شکل ۳- نمودار جریان فرآیند بازسازی داده های گمشده با روش MICE

Figure 3- Flowchart of the missing data imputation process using MICE

این روش با الگوریتم های پیشرفته یادگیری ماشین می تواند دقت و کارایی فرآیند بازسازی داده ها را به طور قابل توجهی بهبود بخشد.

ترکیب روش MICE با الگوریتم های یادگیری ماشین: در سال های اخیر، الگوریتم های یادگیری ماشین به ویژه روش های غیر خطی و مبتنی بر پیش بینی روابط پیچیده، به عنوان ابزارهای قدرتمند در مدل سازی داده های اقلیمی معرفی شده اند (Davari et al., 2025; Ke et al., 2017). از آنجا که MICE یک

مزایای اصلی روش MICE شامل امکان استفاده در داده های چندمتغیره و پیچیده، حفظ همبستگی ها و ساختار وابستگی بین متغیرها، انعطاف پذیری در انتخاب مدل رگرسیونی مناسب برای هر متغیر (خطی، لجستیک و غیره) و قابلیت برآورد عدم قطعیت ناشی از جایگزینی داده ها از طریق تولید چند مجموعه داده تکمیل شده است. با وجود این مزایا، عملکرد MICE در شرایطی که روابط بین متغیرها غیر خطی یا بسیار پیچیده باشند محدود شده و دقت بازسازی کاهش می یابد (Azur et al., 2011). از این رو، ترکیب

آماري و ساختار همبستگی بین متغیرها توسط MICE، دقت بازسازی داده‌ها را در شرایط پیچیده اقلیمی نیز با استفاده از الگوریتم‌های یادگیری ماشین افزایش می‌دهد (Alejo-Sanchez et al., 2025). مطالعات اخیر نیز کارایی چنین رویکردهای ترکیبی را در حوزه اقلیم و هیدرولوژی نشان داده‌اند؛ برای مثال، Farzandi et al., (2022) در ایران گزارش کرده‌اند که روش‌های ترکیبی می‌توانند عملکرد به مراتب بهتری نسبت به روش‌های کلاسیک ارائه دهند.

۲-۴- رویکرد اعتبارسنجی

برای ارزیابی عملکرد مدل‌ها از روش اعتبارسنجی متقابل پنج‌تایی^۱ مبتنی بر سری زمانی^۲ استفاده شد؛ به گونه‌ای که در هر تکرار، داده‌های مربوط به دوره‌های زمانی گذشته برای آموزش و داده‌های دوره بعدی برای آزمون مورد استفاده قرار گرفتند و از اختلاط تصادفی داده‌ها جلوگیری شد (Tashman, 2000). در این روش، داده‌های آموزشی به پنج زیرمجموعه مساوی تقسیم شدند و در هر تکرار، چهار زیرمجموعه برای آموزش مدل و یک زیرمجموعه برای اعتبارسنجی به کار گرفته شد؛ این فرآیند به گونه‌ای تکرار شد که هر زیرمجموعه یک بار به عنوان داده‌های اعتبارسنجی استفاده شود. میانگین نتایج حاصل از پنج مرحله به عنوان معیار نهایی عملکرد مدل گزارش گردید تا برآوردی پایدار و قابل اعتماد از توانایی مدل فراهم شود. برای سنجش دقت بازسازی داده‌های گمشده، مقادیر واقعی حذف شده به طور عمدی به عنوان مرجع مورد استفاده قرار گرفتند و مقادیر بازسازی شده توسط مدل‌ها با این مقادیر واقعی مقایسه شدند. این رویکرد امکان بررسی توانایی مدل‌ها در برآورد دقیق مقادیر واقعی تحت شرایط داده‌های ناقص را فراهم کرده و مزایای کلیدی از جمله کاهش احتمال بیش‌برازش، شبیه‌سازی شرایط واقعی شبکه‌های اقلیم‌سنجی و ایجاد امکان مقایسه مستقیم میان روش‌ها تحت شرایط یکسان را فراهم می‌کند (Géron, 2022).

۲-۴- شاخص‌های ارزیابی

به منظور 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، اریبی و تغییرپذیری را ارزیابی می‌کنند. شاخص‌های مورد استفاده شامل ضریب تعیین (R^2)، خطای جذر میانگین مربعات نرمال شده (NRMSE)، درصد

چارچوب انعطاف‌پذیر مبتنی بر مدل‌سازی شرطی است و نوع مدل پیش‌بینی در هر مرحله قابل تنظیم می‌باشد، در این پژوهش چهار الگوریتم مختلف به عنوان مدل شرطی در ساختار MICE مورد استفاده قرار گرفتند تا امکان مقایسه روابط خطی و غیرخطی فراهم شود.

MICE-LR: استفاده صریح از رگرسیون خطی به عنوان مدل شرطی در چارچوب MICE، با هدف ایجاد مبنای مقایسه با مدل‌های غیرخطی.

MICE-DT: بهره‌گیری از درخت تصمیم برای پیش‌بینی مقادیر گمشده و مدل‌سازی روابط غیرخطی و پیچیده.

MICE-KNN: پیش‌بینی مقادیر گمشده بر اساس نزدیک‌ترین نمونه‌ها در فضای ویژگی‌ها، مناسب داده‌های پراکنده و غیرمتوازن.

MICE-SVM: الگوریتم SVM با کرنل مناسب، توانایی مدل‌سازی روابط غیرخطی پیچیده بین متغیرها را دارد و دقت بازسازی داده‌ها را افزایش می‌دهد.

در چارچوب MICE، ابتدا مقادیر گمشده هر متغیر با استفاده از یک روش مقداردهی اولیه ساده (میانگین یا رگرسیون اولیه) جایگزین شد. سپس، فرآیند بازسازی به صورت تکراری و متغیر به متغیر انجام گرفت. در هر مرحله، یک متغیر به عنوان متغیر هدف در نظر گرفته شده و مقادیر مشاهده شده آن به عنوان داده آموزشی برای برازش مدل مورد نظر (DT، LR، KNN یا SVM) استفاده شد. سایر متغیرها به عنوان متغیرهای پیش‌بینی در مدل وارد گردیدند. پس از آموزش مدل بر روی داده‌های کامل، مقادیر گمشده متغیر هدف پیش‌بینی و جایگزین شدند. این فرآیند برای تمامی متغیرهای دارای داده مفقود تکرار شد و یک چرخه کامل تشکیل داد. کل فرآیند طی چندین تکرار متوالی ادامه یافت تا تغییرات مقادیر برآورد شده به حداقل برسد و همگرایی حاصل شود. در مدل‌های ترکیبی (Hybrid MICE)، تنها تفاوت در نوع مدل شرطی مورد استفاده در هر مرحله است؛ به گونه‌ای که به جای رگرسیون کلاسیک، از الگوریتم‌های یادگیری ماشین به عنوان مدل پیش‌بینی استفاده شد.

ترکیب این الگوریتم‌ها با MICE به گونه‌ای طراحی شده است که مزایای روش آماری بازسازی چندگانه و توانمندی‌های الگوریتم‌های یادگیری ماشین در مدل‌سازی غیرخطی به طور همزمان بهره‌برداری شوند. این ترکیب علاوه بر حفظ ویژگی‌های

² Time-Series Split

¹ 5-fold cross-validation

اریبی و تغییرپذیری را در نظر می گیرد (Gupta Hoshin et al., 1999; Knob et al., 2019).

$$KGE = 1 - \sqrt{(1-r)^2 + (1-\beta)^2 + (1-\gamma)^2} \quad (4)$$

در رابطه ۴: r ضریب همبستگی بین مقادیر پیش بینی و واقعی، $\beta = \frac{\bar{y}_m}{\bar{y}_o}$ نسبت میانگین پیش بینی به میانگین واقعی (نمایگر اریبی) و $\gamma = \frac{CV_m}{CV_o}$ نسبت ضریب تغییرات پیش بینی به واقعی (نمایگر تغییرپذیری) است. مقدار ایده آل شاخص برابر ۱ است و هرچه مقدار آن به ۱ نزدیک تر باشد، عملکرد مدل مطلوب تر خواهد بود.

۳- نتایج و بحث

۳-۱- جمع بندی عملکرد مدل ها در بازسازی داده های گمشده اقلیمی

به منظور ارائه یک دیدگاه کلی از عملکرد مدل های مورد بررسی، نتایج بازسازی داده های گمشده اقلیمی به صورت خلاصه و دسته بندی شده در جدول ۴ ارائه شده است. در این جدول، برای هر دسته از متغیرهای اقلیمی و برای هر ایستگاه، بهترین مقدار هر شاخص ارزیابی (R^2 ، NRMSE و KGE) به همراه مدل متناظر با آن گزارش شده است. این رویکرد امکان مقایسه ی جامع و یکپارچه ی عملکرد مدل ها را در مقیاس حوضه ای فراهم می کند. نتایج کلی جدول ۴ نشان می دهد که عملکرد مدل ها به طور قابل توجهی به نوع متغیر اقلیمی وابسته است و هیچ مدل واحدی برای تمامی متغیرها و ایستگاه ها به عنوان روش برتر مطلق شناسایی نمی شود. با این حال، الگوهای مشخصی در برتری نسبی مدل ها در دسته های مختلف اقلیمی قابل مشاهده است.

در میان تمام دسته ها، متغیرهای دمایی بیشترین دقت بازسازی را نشان می دهند. مقادیر R^2 در تمامی ایستگاه ها بیش از ۰/۹۵ و مقادیر KGE عمدتاً بالاتر از ۰/۹۴ است، که نشان دهنده انطباق بسیار خوب داده های بازسازی شده با مشاهدات واقعی می باشد. در این دسته، مدل MICE-DT در اکثر ایستگاه ها (ارومیه، سراب، سقز و مراغه) بهترین عملکرد را از نظر R^2 و KGE ارائه کرده است، در حالی که مدل MICE-SVM تنها در ایستگاه بناب به عنوان مدل برتر ظاهر شده است. علاوه بر این، مقادیر NRMSE نسبتاً پایین (عموماً کمتر از ۱۵ درصد) نشان می دهد که

اریبی (PBIAS) و کارایی کلینگ-گوپتا (KGE) هستند. در ادامه، رابطه و تفسیر هر یک از شاخص ها ارائه می شود.

• ضریب تعیین (R^2)

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

در رابطه (۱)، y_i و $\hat{y}_i$ به ترتیب نشان دهنده مقادیر مشاهده شده و پیش بینی شده و n تعداد کل مشاهدات است. این شاخص میزان همبستگی خطی بین مقادیر بازسازی شده و مقادیر واقعی را نشان می دهد. مقدار آن در بازه ۰ تا ۱ قرار دارد؛ مقادیر نزدیک به ۱ بیانگر برازش مطلوب تر مدل است (Willmott, 1981; Legates & McCabe Jr, 1999).

• خطای جذر میانگین مربعات نرمال شده ($NRMSE$)

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

با توجه به رابطه (۲)، شاخص NRMSE هنگام تقسیم بر میانگین مقادیر واقعی، همواره مقداری غیرمنفی دارد و هرچه به صفر نزدیک تر باشد، دقت پیش بینی مدل بالاتر است، و مقادیر بزرگ تر نشان دهنده خطای نسبی بیشتر نسبت به میانگین داده ها هستند (Jamali Jeze et al., 2020; Legates & McCabe Jr, 1999).

• درصد اریبی (PBIAS)

$$PBIAS = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)}{\sum_{i=1}^n y_i} \times 100 \quad (3)$$

بر اساس رابطه (۳)، PBIAS میزان تمایل سیستماتیک خطای مدل را نشان می دهد؛ مقدار صفر بیانگر پیش بینی بدون اریبی است، مقادیر مثبت نشان دهنده پیش بینی بیش از حد و مقادیر منفی نشان دهنده پیش بینی کمتر از مقادیر واقعی می باشد، به طوری که PBIAS در حدود ± 10 درصد کیفیت عالی، ± 20 تا ± 10 درصد کیفیت خوب، ± 30 تا ± 20 درصد کیفیت متوسط و مقادیر بیش از ± 30 درصد کیفیت ضعیف مدل را نشان می دهد (Gupta Hoshin et al., 1999; N. Moriasi et al., 2007).

• کارایی کلینگ-گوپتا (KGE)

معیاری جامع برای ارزیابی عملکرد مدل های هیدرولوژیکی است که همزمان سه جنبه مهم پیش بینی مدل یعنی همبستگی،

³ Kling-Gupta Efficiency

¹ Coefficient of Determination

² Normal Root Mean Squared Error

شدید و رفتار پرنوسان باد، وابستگی شدید به شرایط محلی و توپوگرافی، همبستگی پایین آن با سایر متغیرهای اقلیمی، و توزیع آماری نامنظم آن باشد. به همین دلیل، مدل‌های MICE و رویکردهای ترکیبی آن در پیش‌بینی باد نسبت به متغیرهایی مانند دما یا رطوبت با چالش بیشتری مواجه هستند. مقادیر R^2 در اکثر ایستگاه‌ها کمتر از ۰/۵۵ و در برخی موارد نزدیک به ۰/۳۵ است، در حالی که مقادیر NRMSE بالا (حدود ۲۵ تا ۴۰ درصد) بیانگر خطای نسبی زیاد در بازسازی این متغیر می‌باشد. بر اساس جدول ۴، هیچ مدل غالبی برای بازسازی باد مشاهده نمی‌شود. اگرچه MICE-SVM در ایستگاه بناب عملکرد بهتری نشان داده است، اما در سایر ایستگاه‌ها روش‌های ساده‌تر مانند MICE یا MICE-LR نیز به‌عنوان بهترین مدل ظاهر شده‌اند. این پراکندگی عملکرد بیانگر ماهیت بسیار ناپایدار، محلی و آشوبناک داده‌های باد است. در بازسازی متغیرهای رطوبت و بخار آب، عملکرد مدل‌ها نسبت به ابرناکی و باد بهبود یافته است. مقادیر R^2 عمدتاً بالاتر از ۰/۹۰ و مقادیر NRMSE در اغلب ایستگاه‌ها کمتر از ۱۰ درصد است که نشان‌دهنده دقت مناسب بازسازی این متغیرها می‌باشد.

خطای نسبی بازسازی دما محدود و قابل قبول است. این نتایج بیانگر آن است که روابط بین متغیرهای دمایی و سایر پارامترهای اقلیمی عمدتاً پایدار و قابل پیش‌بینی هستند و مدل‌های مبتنی بر درخت تصمیم توانایی بالایی در استخراج این روابط دارند.

برای متغیرهای ابرناکی، دقت بازسازی به‌طور محسوسی کاهش یافته است. مقادیر R^2 در بازه ۰/۷۸ تا ۰/۸۶ و مقادیر NRMSE عموماً بزرگ‌تر از ۳۰ درصد هستند که نشان‌دهنده نوسانات زیاد و ماهیت گسسته این متغیر است. با این حال، الگوی غالب جدول ۴ نشان می‌دهد که مدل MICE-DT در اغلب ایستگاه‌ها بهترین عملکرد را بر اساس شاخص‌های R^2 و KGE داشته است، در حالی که MICE-SVM تنها در ایستگاه بناب عملکرد بهتری ارائه کرده است. این برتری درخت تصمیم نشان می‌دهد که تغییرات ابرناکی عمدتاً تحت تأثیر آستانه‌ها و قواعد شرطی قرار دارند و کمتر از روابط پیوسته و خطی تبعیت می‌کنند. نتایج نشان داد که بازسازی سرعت باد نسبت به سایر متغیرهای اقلیمی ضعیف‌تر بود. این موضوع می‌تواند به دلیل نوسانات روزانه

جدول ۴- جمع‌بندی عملکرد برتر مدل‌های بازسازی داده‌های گمشده اقلیمی در ایستگاه‌های منتخب حوضه دریاچه ارومیه

Table 4- Summary of the best-performing climate data imputation models across selected stations in the Urmia Lake basin

Category	Index	Urmia	Bonab	Tabriz	Sarab	Saghez	Maragheh
Temperature	R^2	0.979 (MICE-DT)	0.975 (MICE-SVM)	0.976 (MICE)	0.963 (MICE-DT)	0.953 (MICE-DT)	0.971 (MICE-DT)
	NRMSE	18.182 (MICE-DT)	6.740 (MICE-SVM)	8.066 (MICE)	12.393 (MICE)	14.218 (MICE-DT)	10.455 (MICE)
	KGE	0.985 (MICE-DT)	0.979 (MICE-SVM)	0.975 (MICE-DT)	0.960 (MICE-DT)	0.941 (MICE-DT)	0.975 (MICE-DT)
Cloud Cover	R^2	0.797 (MICE-DT)	0.856 (MICE-SVM)	0.798 (MICE)	0.782 (MICE-DT)	0.826 (MICE-DT)	0.796 (MICE-LR)
	NRMSE	34.73 (MICE-DT)	36.107 (MICE-SVM)	30.140 (MICE)	33.389 (MICE)	43.622 (MICE-DT)	37.650 (MICE-LR)
	KGE	0.882 (MICE-DT)	0.914 (MICE-SVM)	0.872 (MICE)	0.867 (MICE-DT)	0.894 (MICE-DT)	0.875 (MICE-LR)
Wind	R^2	0.370 (MICE)	0.597 (MICE-SVM)	0.550 (MICE)	0.540 (MICE)	0.342 (MICE)	0.564 (MICE)
	NRMSE	39.145 (MICE)	38.000 (MICE-SVM)	25.354 (MICE)	29.741 (MICE)	31.604 (MICE)	27.693 (MICE)
	KGE	0.542 (MICE)	0.756 (MICE-SVM)	0.695 (MICE-LR)	0.678 (MICE-LR)	0.474 (MICE-LR)	0.699 (MICE-LR)
Humidity & Vapor	R^2	0.913 (MICE)	0.953 (MICE-SVM)	0.929 (MICE)	0.888 (MICE)	0.929 (MICE)	0.942 (MICE)
	NRMSE	8.960 (MICE-DT)	6.605 (MICE-SVM)	7.982 (MICE)	8.689 (MICE)	9.910 (MICE)	9.055 (MICE-LR)
	KGE	0.946 ^a (MICE)	0.965 (MICE-SVM)	0.945 (MICE-LR)	0.928 (MICE-LR)	0.949 (MICE-LR)	0.963 (MICE-LR)
Radiation & Energy	R^2	0.784 (MICE)	0.809 (MICE-SVM)	0.732 (MICE)	0.693 (MICE)	0.774 (MICE)	0.755 (MICE)
	NRMSE	22.050 (MICE)	18.168 (MICE-SVM)	21.516 (MICE)	21.298 (MICE)	18.192 (MICE)	20.825 (MICE)
	KGE	0.872 (MICE)	0.869 (MICE-SVM)	0.808 (MICE)	0.796 (MICE)	0.825 (MICE-LR)	0.833 (MICE-LR)
Pressure	R^2	0.327 (MICE-DT)	0.430 (MICE-SVM)	0.396 (MICE)	0.279 (MICE)	0.423 (MICE)	0.387 (MICE)
	NRMSE	0.413 (MICE-DT)	0.342 (MICE-SVM)	0.340 (MICE)	0.327 (MICE)	0.312 (MICE)	0.643 (MICE)
	KGE	0.502 (MICE-DT)	0.570 (MICE-SVM)	0.537 (MICE-DT)	0.421 (MICE-DT)	0.544 (MICE-DT)	0.556 (MICE-LR)

الگوریتم‌های یادگیری ماشین می‌تواند، در صورت انتخاب آگاهانه مدل، دقت بازسازی را در مطالعات اقلیم‌شناسی و مدیریت منابع آب به‌طور قابل توجهی افزایش دهد.

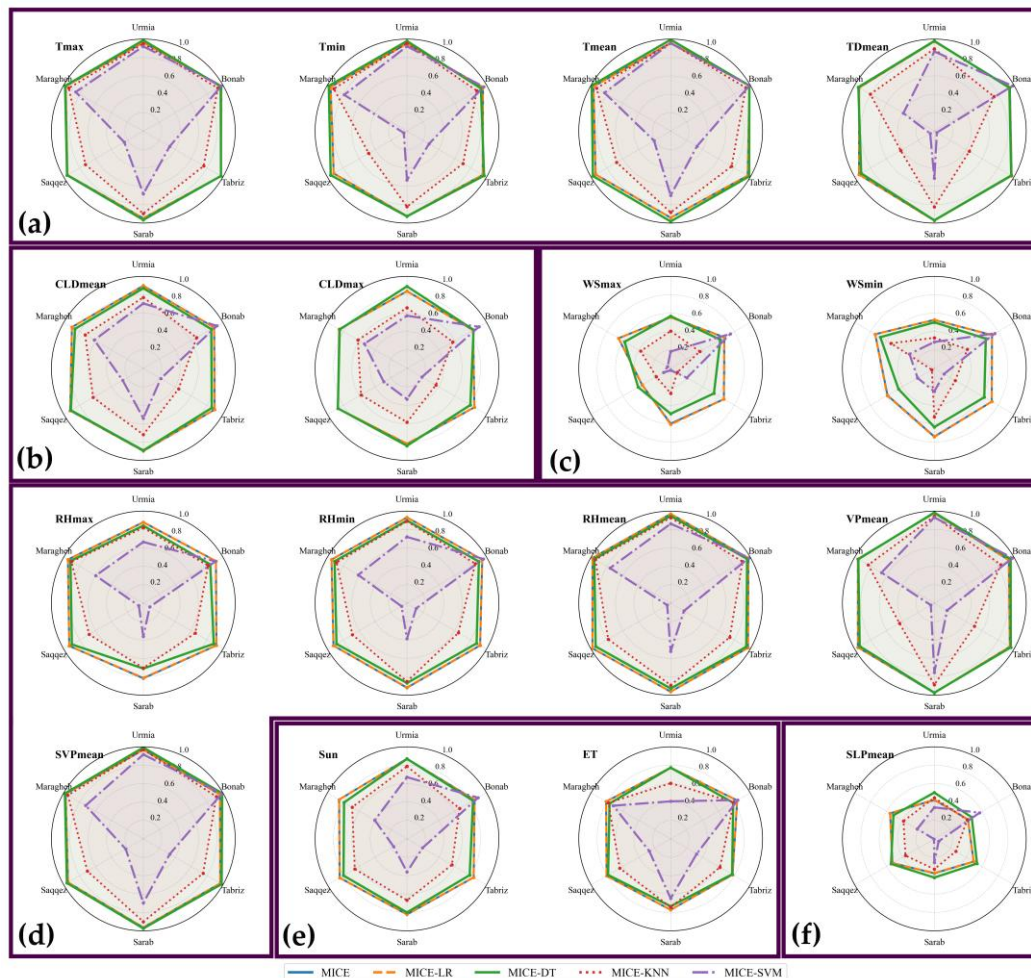
۳-۲- تحلیل مقایسه‌ای عملکرد مدل‌ها به تفکیک متغیر اقلیمی

در این بخش، عملکرد مدل‌های مختلف بازسازی داده‌های گمشده در چارچوب MICE به تفکیک گروه‌های اصلی متغیرهای اقلیمی مورد بررسی قرار گرفت. شکل ۴، مقایسه‌ی بصری عملکرد مدل‌ها را بر اساس شاخص KGE برای شش گروه متغیر اقلیمی شامل دما (a)، ابرناکی (b)، باد (c)، رطوبت و بخار آب (d)، تابش و انرژی (e) و فشار سطح دریا (f) را در ایستگاه‌های منتخب حوضه دریاچه ارومیه نشان می‌دهد. نتایج نشان می‌دهد که عملکرد مدل‌ها بسته به گروه متغیر و ایستگاه متفاوت است؛ به‌طوری‌که مدل MICE-SVM در ایستگاه بناب و مدل MICE-LR در ایستگاه‌های سراب، سقز و مراغه برتری داشته‌اند، در حالی که روش پایه MICE در برخی ایستگاه‌ها همچنان عملکرد مناسبی داشت. به‌طور کلی، متغیرهای دمایی بالاترین دقت بازسازی را نشان می‌دهند، در حالی که بازسازی متغیرهای بادی دشوارتر بوده و کمترین KGE را به خود اختصاص داده است که این امر ناشی از نوسانات بالای باد، رفتار پرنوسان و همبستگی پایین آن با سایر متغیرهای اقلیمی است. این تحلیل تصویری، علاوه بر داده‌های عددی ارائه شده در جداول پیشین، دید روشنی از توانایی نسبی مدل‌ها در بازسازی داده‌های گمشده فراهم می‌کند و امکان شناسایی الگوهای مکانی عملکرد مدل‌ها را نیز میسر می‌سازد.

۳-۲-۱- متغیرهای دمایی

نتایج مربوط به متغیرهای دمایی (Tmax، Tmin، Tmean و TDmean) نشان می‌دهد که تمامی مدل‌ها در بازسازی داده‌های دما عملکرد نسبتاً بالایی داشته‌اند؛ با این حال، تفاوت‌های معناداری بین مدل‌ها مشاهده می‌شود (شکل ۴ (a)). مدل MICE-DT در اغلب ایستگاه‌ها و برای تمامی زیرمتغیرهای دمایی، بالاترین مقادیر KGE را ثبت کرده است، به‌ویژه برای دمای میانگین (Tmean) که مقادیر KGE نزدیک به ۰/۹۸ در اکثر ایستگاه‌ها به‌دست آمده است. این موضوع بیان‌گر توانایی بالای درخت تصمیم در مدل‌سازی روابط غیرخطی و وابستگی‌های پیچیده بین متغیرهای دمایی است. در مقابل، مدل‌های MICE-KNN و MICE-SVM نوسانات قابل توجهی در عملکرد خود نشان داده‌اند؛ به‌طوری‌که در برخی ایستگاه‌ها (مانند تبریز و سقز) کاهش شدید مقدار KGE مشاهده می‌شود.

مطالعه نتایج جدول ۴ نشان می‌دهد که مدل MICE-SVM در ایستگاه بناب و مدل MICE-LR در ایستگاه‌های سراب، سقز و مراغه عملکرد برتری داشته‌اند، با این حال روش پایه MICE در برخی ایستگاه‌ها همچنان توانسته عملکرد قابل قبولی ارائه دهد. این موضوع نشان می‌دهد که روابط بین متغیرهای رطوبتی، هم شامل اجزای خطی و هم غیرخطی است. برای متغیرهای تابش و انرژی، مقادیر R^2 در سطح متوسط (حدود ۰/۶۹ تا ۰/۸۱) قرار دارند و روش پایه MICE در اکثر ایستگاه‌ها به‌عنوان بهترین مدل شناسایی شده است. این نتیجه نشان می‌دهد که برای این متغیرها، حفظ ساختار همبستگی چندمتغیره اهمیت بیشتری نسبت به استفاده از مدل‌های پیچیده‌تر دارد. در مورد متغیر فشار سطح دریا، دقت بازسازی در پایین‌ترین سطح قرار دارد. مقادیر R^2 عمدتاً کمتر از ۰/۴۳ است، اما مدل MICE-DT بر اساس شاخص KGE در اکثر ایستگاه‌ها عملکرد بهتری داشته است که نشان‌دهنده توان این مدل در بازسازی الگوی نوسانات فشار، حتی در صورت دامنه تغییرات محدود، می‌باشد. نتایج این پژوهش نشان می‌دهد که کارایی مدل‌های بازسازی داده‌های گمشده اقلیمی به‌طور مستقیم تحت تأثیر پیچیدگی فیزیکی و آماری متغیر مورد بررسی قرار دارد. این یافته با نتایج Diouf et al. (2025) و Jing et al. (2022) هم‌راستا است که گزارش کردند متغیرهای دارای همبستگی بین‌متغیره قوی و پیوستگی زمانی بالا، با دقت بیشتری توسط روش‌های چندمتغیره بازسازی می‌شوند. در پژوهش حاضر نیز متغیرهایی مانند دما و رطوبت که دارای پیوستگی زمانی بالا و روابط فیزیکی قوی با سایر پارامترها هستند، با دقت بالاتری بازسازی شده‌اند. در مقابل، متغیرهایی نظیر باد، ابرناکی و فشار که رفتار ناپایدارتر و وابستگی محلی بیشتری دارند، چالش‌برانگیزتر بوده‌اند؛ موضوعی که در مطالعات Kumar and Dwarakish (2025) نیز به آن اشاره شده است. برتری مدل MICE-DT در متغیرهای دمایی، ابرناکی و فشار نشان می‌دهد که ساختار درخت تصمیم توانایی بالایی در استخراج قواعد شرطی و آستانه‌ای از داده‌های اقلیمی دارد. از سوی دیگر، عملکرد مناسب MICE-SVM در متغیرهای رطوبتی بیانگر اهمیت مدل‌سازی روابط غیرخطی و پیچیده در فرآیندهای ترمودینامیکی جو است. نکته قابل توجه این است که روش پایه MICE در برخی متغیرها مانند تابش و انرژی عملکردی قابل رقابت با مدل‌های ترکیبی ارائه داده است. این موضوع نشان می‌دهد که افزایش پیچیدگی مدل لزوماً منجر به بهبود عملکرد نمی‌شود و انتخاب مدل باید متناسب با ماهیت متغیر صورت گیرد. در مجموع، نتایج این مطالعه بر ضرورت استفاده از رویکردهای تطبیقی و متغیرمحور در بازسازی داده‌های گمشده اقلیمی تأکید می‌کند و نشان می‌دهد که ترکیب MICE با



شکل ۴- مقایسه عملکرد مدل‌های بازسازی بر اساس شاخص KGE برای شش گروه متغیر اقلیمی شامل دما (a)، ابرناکی (b)، باد (c)، رطوبت و بخار آب (d)، تابش و انرژی (e) و فشار سطح دریا (f)

Figure 4. Performance comparison of reconstruction models based on the KGE index for six groups of climate variables including temperature (a), cloudiness (b), wind (c), humidity and water vapor (d), radiation and energy (e), and sea-level pressure (f)

متغیر گسسته-نیمه پیوسته با رفتار تصادفی‌تر بوده و وابستگی خطی ضعیف‌تری با سایر متغیرهای اقلیمی دارد (Wilks, D.S., 2011).

۳-۲-۳- متغیرهای باد

نتایج مربوط به WSmax و WSmin نشان‌دهنده چالش برانگیز بودن بازسازی داده‌های باد است (شکل ۴ (c)). مقادیر KGE در تمامی مدل‌ها به‌طور محسوسی کمتر از دما و رطوبت بوده و پراکندگی بین ایستگاه‌ها افزایش یافته است. در این گروه، مدل‌های MICE و MICE-LR عملکرد نسبتاً پایدارتری داشته‌اند، در حالی که MICE-DT به‌طور متوسط عملکرد ضعیف‌تری نسبت به دما از خود نشان داده است. عملکرد بسیار ضعیف مدل MICE-SVM در برخی ایستگاه‌ها (مقادیر KGE نزدیک به صفر) بیان‌گر عدم توانایی این مدل در بازسازی نوسانات شدید و رفتار آشفته سرعت باد است. این نتایج نشان می‌دهد که متغیرهای باد به دلیل ماهیت تصادفی، وابستگی ضعیف مکانی و

این ضعف می‌تواند ناشی از حساسیت این الگوریتم‌ها به توزیع داده‌ها، ناهمگنی اقلیمی منطقه و یا محدودیت در تعمیم‌پذیری روابط محلی باشد. به‌طور کلی، نتایج این بخش نشان می‌دهد که متغیرهای دمایی به دلیل پیوستگی زمانی بالا و همبستگی قوی با سایر متغیرها، بیشترین سازگاری را با چارچوب MICE دارند.

۳-۲-۳- متغیرهای ابرناکی

با توجه به شکل ۴ (b) مشاهده شد که برای متغیرهای CLDmax و CLDmean، عملکرد مدل‌ها نسبت به متغیرهای دمایی کاهش یافته است، اما همچنان مدل‌های مبتنی بر MICE و MICE-LR عملکرد پایدار و قابل قبولی را در اغلب ایستگاه‌ها حفظ کرده‌اند. مدل MICE-DT در برخی ایستگاه‌ها عملکرد مناسبی دارد، اما یکنواختی آن نسبت به دما کمتر است. در مقابل، مدل‌های MICE-SVM و MICE-KNN مجدداً عملکرد ضعیف‌تری، به‌ویژه در ایستگاه‌هایی با تغییرپذیری بالای ابرناکی نشان می‌دهند. این موضوع قابل انتظار است، زیرا ابرناکی یک

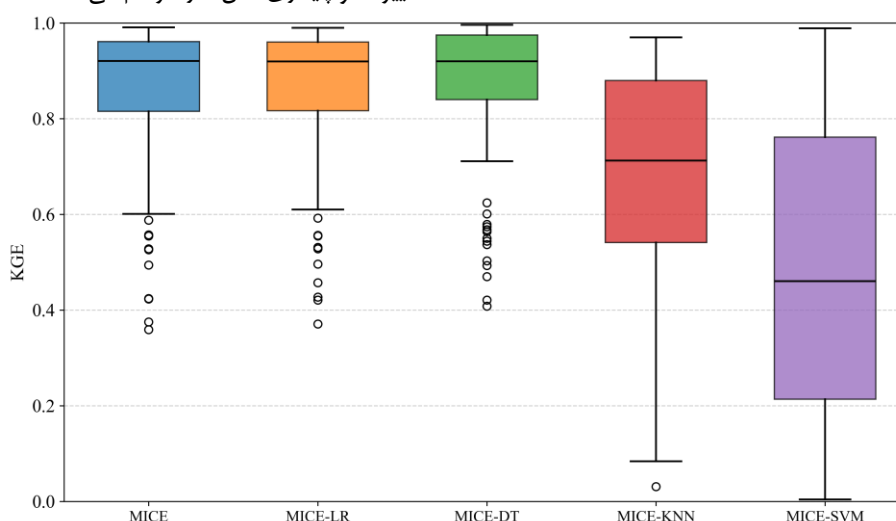
۳-۲-۶- فشار سطح دریا

نتایج مربوط به SLPmean نشان می‌دهد که این متغیر کمترین مقدار KGE را در میان تمامی متغیرها دارد. اگرچه فشار سطح دریا از نظر فیزیکی دارای تغییرپذیری کمتری نسبت به سایر متغیرهاست، اما بازسازی نوسانات کوچک آن بسیار حساس بوده و خطاهای نسبی منجر به افت شاخص KGE می‌شود. در این گروه، مدل MICE-DT نسبت به سایر مدل‌ها عملکرد بهتری داشته است، در حالی که MICE-SVM ضعیف‌ترین عملکرد را ثبت کرده است. این نتایج نشان می‌دهد که حتی برای متغیرهای نسبتاً پایدار، انتخاب مدل مناسب نقش کلیدی در کیفیت بازسازی داده‌های گمشده ایفا می‌کند.

به‌طور کلی، نتایج این بخش نشان می‌دهد که کارایی مدل‌های بازسازی داده‌های گمشده به‌شدت تابع نوع متغیر اقلیمی است. مدل MICE-DT در اغلب گروه‌های متغیری، به‌ویژه دما و رطوبت، بهترین عملکرد را ارائه داده است، در حالی که مدل‌های مبتنی بر KNN و SVM ناپایداری بیشتری از خود نشان داده‌اند. این یافته‌ها اهمیت انتخاب مدل متناسب با ماهیت فیزیکی متغیر و ساختار داده‌ها را برجسته می‌کند.

۳-۳- تحلیل مدل‌محور عملکرد روش‌های بازسازی داده‌های گمشده

به‌منظور مقایسه جامع عملکرد مدل‌های بازسازی داده‌های گمشده، نمودار جعبه‌ای شاخص KGE برای پنج مدل MICE-، MICE-LR، MICE-DT، MICE-KNN و MICE-SVM با تجمیع نتایج تمامی متغیرهای اقلیمی و ایستگاه‌ها ترسیم شد (شکل ۵). این نمودار امکان بررسی هم‌زمان سطح عملکرد مرکزی، دامنه تغییرات و پایداری مدل‌ها را فراهم می‌کند.



شکل ۵- مقایسه عملکرد مدل‌ها در بازسازی داده‌های گمشده با استفاده از شاخص KGE

Figure 5. Comparison of model performance in missing data imputation using KGE

حساسیت به شرایط محلی، کمترین سازگاری را با چارچوب‌های یادگیری ماشین دارند.

۳-۲-۴- متغیرهای رطوبت و بخار آب

شکل ۴ (d) نتایج بازسازی داده‌های گمشده متغیرهای رطوبت و بخار آب را نشان داده است. برای متغیرهای RHmax، RHmean، RHmin و VPmean، نتایج حاکی از عملکرد بسیار مطلوب مدل‌های MICE، MICE-LR و MICE-DT است. در بسیاری از ایستگاه‌ها، مقادیر KGE بالاتر از ۰/۹۵ مشاهده می‌شود که نشان‌دهنده دقت بالای بازسازی این متغیرها است. این عملکرد مناسب را می‌توان به همبستگی قوی بین رطوبت، دما و فشار بخار نسبت داد که امکان بهره‌گیری مؤثر از ساختار چندمتغیره MICE را فراهم می‌کند. در مقابل، مدل‌های MICE-KNN و MICE-SVM همچنان ناپایداری بیشتری نشان داده‌اند، اگرچه عملکرد آن‌ها نسبت به متغیرهای باد بهتر است.

۳-۲-۵- متغیرهای تابش و انرژی

در متغیرهای Sun و ET، عملکرد مدل‌ها در سطح متوسط ارزیابی می‌شود (شکل ۴ (e)). مدل‌های MICE و MICE-LR بهترین عملکرد را در اغلب ایستگاه‌ها ارائه داده‌اند، در حالی که MICE-DT در برخی ایستگاه‌ها افت عملکرد نشان می‌دهد. تابش خورشیدی و تبخیر-تعرق به‌شدت تحت تأثیر عوامل محلی، پوشش ابری و شرایط سینوپتیکی هستند که این امر بازسازی دقیق آن‌ها را دشوارتر می‌کند. عملکرد ضعیف‌تر MICE-SVM در این گروه نیز بیانگر محدودیت این مدل در استخراج روابط پایدار میان متغیرهای انرژی-تابشی است.

MICE-KNN و MICE-SVM می‌تواند ناشی از حساسیت این الگوریتم‌ها به تراکم داده، مقیاس متغیرها و ناهمگنی فضایی ایستگاه‌ها باشد. به‌ویژه در شرایطی که توزیع داده‌ها نامتوازن یا دارای نویز بالا است، این مدل‌ها قادر به تعمیم مناسب الگوها در تمامی متغیرها نبوده‌اند. به‌طور کلی، تحلیل مدل‌محور نشان می‌دهد که مدل‌های مبتنی بر درخت تصمیم و ساختارهای ساده‌تر آماری نسبت به روش‌های مبتنی بر فاصله یا کرنل، از پایداری و قابلیت تعمیم بالاتری در بازسازی داده‌های گمشده اقلیمی برخوردارند. این یافته اهمیت انتخاب مدل مناسب را متناسب با ماهیت داده‌های اقلیمی و هدف مطالعه برجسته می‌سازد.

۳-۴- تحلیل ایستگاه‌محور عملکرد مدل‌ها

بر اساس جدول ۵ و مقایسه KGE برای شش ایستگاه منتخب نشان‌دهنده عملکرد متفاوت مدل‌های بازسازی داده‌های گمشده است و نکات مهمی در زمینه دقت و پایداری مدل‌ها ارائه می‌دهد. مشاهده می‌شود که مدل‌های MICE و MICE-LR به‌طور کلی در تمام ایستگاه‌ها عملکردی پایدار و مشابه دارند، به‌طوری که KGE آنها معمولاً در محدوده بالا و نزدیک به مقدار ۱ قرار دارد. این امر بیان‌گر توانایی این مدل‌ها در حفظ روند کلی داده‌های اقلیمی و حداقل خطا در بازسازی مقادیر گمشده است. از سوی دیگر، مدل MICE-DT با ارائه مقادیر KGE بالاتر در اکثر ایستگاه‌ها، توانسته است دقت بازسازی را افزایش دهد و بازسازی داده‌ها را به شکل قابل اعتمادتری انجام دهد.

نتایج نشان می‌دهد که مدل MICE-DT بالاترین مقدار میانگین KGE را در بین تمامی مدل‌ها داراست و دامنه بین چارکی آن نیز نسبتاً محدود است که بیانگر عملکرد قوی و پایدار این مدل در بازسازی داده‌های گمشده اقلیمی می‌باشد. پس از آن، مدل‌های MICE و MICE-LR عملکردی بسیار نزدیک به یکدیگر از خود نشان داده‌اند؛ به‌طوری که میانگین KGE این دو مدل در محدوده مقادیر بالا قرار داشته و پراکندگی نتایج آن‌ها محدود است. در مقابل، مدل MICE-KNN با میانگین پایین‌تر و دامنه تغییرات وسیع‌تر، ناپایداری بیشتری را در بازسازی متغیرهای مختلف نشان می‌دهد. این وضعیت در مورد مدل MICE-SVM تشدید شده است؛ به‌طوری که این مدل علاوه بر میانگین پایین‌تر، دارای داده‌های پرت قابل توجه و پراکندگی زیاد در مقادیر KGE می‌باشد که حاکی از حساسیت بالای عملکرد آن به نوع متغیر و ایستگاه است. برتری مدل MICE-DT را می‌توان به توانایی بالای درخت تصمیم در مدل‌سازی روابط غیرخطی و آستانه‌ای بین متغیرهای اقلیمی نسبت داد. این ویژگی به‌ویژه در داده‌های هواشناسی که اغلب تحت تأثیر فرآیندهای پیچیده فیزیکی قرار دارند، نقش کلیدی ایفا می‌کند. علاوه بر این، ادغام درخت تصمیم در چارچوب MICE باعث شده است که عدم قطعیت ناشی از داده‌های گمشده به‌صورت مؤثرتری مدیریت شود. عملکرد مشابه مدل‌های MICE و MICE-LR نشان می‌دهد که در بسیاری از متغیرهای اقلیمی، روابط خطی یا شبه‌خطی غالب بوده و افزودن ساختار رگرسیون خطی به MICE به بهبود چشمگیری نسبت به نسخه پایه منجر نشده است. در مقابل، عملکرد ضعیف‌تر و ناپایدار مدل‌های

جدول ۵- مقادیر میانگین شاخص KGE برای شش ایستگاه منتخب بر اساس مدل‌های بازسازی داده‌های گمشده

Station	MICE	MICE-LR	MICE-DT	MICE-KNN	MICE-SVM
Urmia	0.915	0.915	0.935	0.780	0.735
Bonab	0.924	0.923	0.940	0.775	0.915
Tabriz	0.909	0.909	0.928	0.609	0.314
Sarab	0.911	0.911	0.925	0.748	0.480
Saqqez	0.879	0.879	0.914	0.511	0.148
Maragheh	0.914	0.914	0.948	0.771	0.636

ارومیه، بناب و مراغه، اکثر مدل‌ها توانسته‌اند مقادیر KGE بالایی ارائه دهند و بازسازی داده‌ها با دقت بالایی انجام شده است. این امر بیان‌گر آن است که تغییرات جوی کم‌تر و ویژگی‌های اقلیمی یکنواخت، باعث افزایش پایداری و دقت مدل‌ها می‌شود. در مقابل، ایستگاه‌هایی با ارتفاع بالاتر و پراکندگی گسترده‌تر داده‌ها مانند سراب، تبریز و سقز، چالش بیش‌تری برای مدل‌ها ایجاد کرده است و باعث کاهش دقت مدل‌هایی شده است که نسبت به ویژگی‌های محلی حساس هستند. نتایج این تحلیل نشان می‌دهد که برای بازسازی داده‌های گمشده در حوزه‌های اقلیمی با تنوع مکانی، استفاده از مدل‌هایی مانند MICE-DT و MICE-LR توصیه

در مقابل، مدل‌های MICE-KNN و MICE-SVM در برخی ایستگاه‌ها، به‌ویژه در ایستگاه‌هایی با تغییرات مکانی و ارتفاعی بیش‌تر مانند تبریز و سقز، دچار کاهش قابل توجه KGE شده‌اند. این موضوع نشان‌دهنده حساسیت این مدل‌ها به پراکندگی داده‌ها و ویژگی‌های محلی است و بیانگر این نکته است که این مدل‌ها در مناطق با تغییر اقلیم شدید و ناهمگونی بیش‌تر، عملکرد پایین‌تری دارند. این تفاوت عملکرد مدل‌ها می‌تواند به تفاوت در روش‌های الگوریتمی آن‌ها و نحوه تطبیق با ویژگی‌های محلی داده‌ها نسبت داده شود. تحلیل مکانی نشان می‌دهد که در ایستگاه‌هایی با ارتفاع متوسط و شرایط اقلیمی نسبتاً پایدار مانند

۳-۵- ارزیابی سوگیری و کارایی محاسباتی مدل های بازسازی داده های گمشده

به منظور تکمیل ارزیابی عملکرد مدل های بازسازی داده های گمشده، در این بخش سوگیری سیستماتیک مدل ها و کارایی محاسباتی آن ها مورد بررسی قرار گرفت. جدول ۶، میانگین قدر مطلق شاخص PBIAS و میانگین زمان اجرای هر مدل را با تجمیع نتایج تمامی متغیرهای اقلیمی و ایستگاه ها ارائه می دهد.

می شود، زیرا این مدل ها توانایی تطبیق بهتر با ویژگی های ایستگاه ها و ارائه بازسازی با دقت بالا را دارند. همچنین، مدل های که حساسیت بیش تری به پراکندگی داده ها دارند، مانند MICE-KNN و MICE-SVM، نیازمند پیش پردازش دقیق تر داده ها یا تنظیمات محلی برای هر ایستگاه هستند تا عملکرد بهینه خود را نشان دهند. به طور کلی، این یافته ها اهمیت درک تفاوت های ایستگاهی و ویژگی های اقلیمی را در انتخاب و کاربرد مدل های بازسازی داده های گمشده برجسته می کنند و پایه ای محکم برای تحلیل های ایستگاه محور و تصمیم گیری در مدیریت داده های اقلیمی فراهم می آورند.

جدول ۶- میانگین قدر مطلق سوگیری (|PBIAS|) و زمان اجرای مدل های بازسازی داده های گمشده

Table 6. Mean absolute PBIAS and computational time of missing data imputation models

Model	PBIAS (%) Mean	Mean runtime (s)
MICE	0.005	9.917
MICE-LR	0.005	8.167
MICE-DT	0.004	11.150
MICE-KNN	0.014	17.350
MICE-SVM	0.033	215.217

استخراج کنند. ساختار شاخه ای درخت تصمیم امکان تفکیک رژیم های اقلیمی مختلف و مدل سازی رفتارهای ناپیوسته را فراهم می کند که می تواند دلیل عملکرد پایدار آن در متغیرهای دمایی و فشاری باشد. همچنین، عملکرد قابل رقابت روش های ساده تر مانند MICE و MICE-LR در برخی متغیرها نشان می دهد که در شرایطی که روابط بین متغیرها عمدتاً خطی و پایدار باشند، استفاده از مدل های پیچیده تر لزوماً منجر به بهبود معنی دار دقت نمی شود؛ نتیجه ای که در مطالعات (Kumar and Dwarakish, 2025) نیز مورد تأکید قرار گرفته است. بنابراین، یافته های این پژوهش تأیید می کند که انتخاب مدل مناسب در بازسازی داده های گمشده باید بر اساس ماهیت آماری و فیزیکی متغیر، سطح داده های مفقود و نیز محدودیت های محاسباتی صورت گیرد. این رویکرد تطبیقی با توصیه های مطرح شده در ادبیات اخیر حوزه مدل سازی اقلیمی همسو است و می تواند در مطالعات اقلیم شناسی و مدیریت منابع آب کاربرد عملی داشته باشد.

۴- نتیجه گیری

در این پژوهش، کارایی مدل های مختلف بازسازی داده های گمشده اقلیمی در چارچوب روش MICE برای داده های روزانه شش ایستگاه منتخب حوضه دریاچه ارومیه مورد ارزیابی قرار گرفت. بدین منظور، روش پایه MICE و چهار مدل ترکیبی مبتنی بر یادگیری ماشین شامل MICE-LR، MICE-DT، MICE-KNN و MICE-SVM برای بازسازی شش گروه

نتایج نشان می دهد که مدل های MICE، MICE-LR و MICE-DT دارای کمترین مقدار |PBIAS| هستند (کمتر از ۰/۰۰۵)، که بیان گر نبود سوگیری سیستماتیک معنادار در بازسازی داده های گمشده است. در مقابل، مدل های MICE-KNN و به ویژه MICE-SVM مقادیر بالاتری از |PBIAS| را نشان می دهند که حاکی از تمایل این مدل ها به بیش برآورد یا کم برآورد سیستماتیک در برخی متغیرها و ایستگاه ها است. از منظر کارایی محاسباتی، اختلاف چشمگیری بین مدل ها مشاهده می شود. مدل های MICE و MICE-LR کمترین زمان اجرا را داشته و از نظر محاسباتی بسیار مقرون به صرفه هستند. مدل MICE-DT با وجود زمان اجرای کمی بالاتر، همچنان در محدوده قابل قبول قرار دارد و با توجه به دقت بالای آن، توازن مناسبی بین دقت و هزینه محاسباتی ارائه می دهد. در مقابل، مدل MICE-SVM با زمان اجرای بسیار بالا (بیش از ۲۰۰ ثانیه به طور میانگین)، کمترین کارایی محاسباتی را داشته و استفاده از آن در مطالعات با حجم داده بالا یا کاربردهای عملیاتی را با محدودیت جدی مواجه می سازد. در مجموع، نتایج این پژوهش نشان می دهد که مدل MICE-DT از نظر همزمان دقت بالا، سوگیری اندک و هزینه محاسباتی معقول، گزینه ای متوازن برای بازسازی داده های گمشده اقلیمی محسوب می شود. این یافته با نتایج مطالعات (Diouf et al., 2023) و (Jing et al., 2022) همخوانی دارد که گزارش کردند مدل های مبتنی بر درخت تصمیم در چارچوب های چندمتغیره قادرند روابط غیرخطی و آستانه ای موجود در داده های اقلیمی را به خوبی

هواشناسی در مطالعات اقلیم‌شناسی و برنامه‌ریزی منابع آب فراهم آورد.

سپاسگزاری

بدینوسیله از حمایت‌های دانشگاه صنعتی اصفهان که سهم بسزایی در موفقیت این پروژه تحقیقاتی داشته‌اند صمیمانه تشکر و قدردانی می‌نماییم.

تضاد منافع نویسندگان

نویسندگان این مقاله اعلام می‌دارند که هیچ‌گونه تضاد منافی در خصوص نگارش و انتشار مطالب و نتایج این پژوهش ندارند.

دسترسی به داده‌ها

مجموعه داده‌های مورد استفاده و/یا تجزیه و تحلیل‌شده در طول مطالعه جاری در صورت درخواست معقول از نویسنده مربوطه در دسترس است.

مشارکت نویسندگان

محمد شایان‌نژاد: نظارت، اعتبارسنجی روش‌شناسی، تفسیر نتایج، بررسی و ویرایش؛ **محمد جمالی:** مفهوم‌سازی، پردازش داده‌ها، مدل‌سازی، تحلیل نرم‌افزار، نگارش - پیش‌نویس اولیه.

منابع

ستاری، م. ت. شیرینی، ک. و جاویدان، س. (۱۴۰۳). ارزیابی کارایی روش‌های کاهش بعد در بهبود دقت مدل‌سازی شاخص کیفی آب با استفاده از الگوریتم‌های یادگیری ماشین. مدل‌سازی مدیریت آب و خاک، ۴(۲)، ۸۹-۱۰۴. doi: 10.22098/mmws.2023.12434.1241
محتشم، جاوید. ۱۴۰۲. بازسازی داده‌های بارندگی روزانه دارای دوره‌های مفقودی، با استفاده از روش‌های محاسبه چندگانه با تطبیق میانگین پیش‌بینی و جنگل تصادفی. پایان‌نامه کارشناسی ارشد، دانشگاه صنعتی اصفهان. ۲۷۸ ص.
وفائی‌ممقانی، آ. اسدی، ا. دربندی، ص. و ستاری، م. ت. (۱۴۰۵). کاربست و مقایسه روش‌های درون‌یابی داده‌های گمشده‌ی تراز آب زیرزمینی با تأکید بر عملکرد DeepMVI (منطقه مورد مطالعه: دشت عجبشیر). مدلسازی مدیریت آب و خاک، ۶(۱)، ۱۷-۳۶. doi: 10.22098/mmws.2025.17457.1601

اصلی متغیرهای اقلیمی به کار گرفته شدند و عملکرد آن‌ها با استفاده از شاخص‌های آماری R^2 ، NRMSE، KGE و PBIAS و همچنین زمان اجرای مدل‌ها مورد مقایسه قرار گرفت. این پژوهش نشان داد که کارایی روش‌های بازسازی داده‌های گمشده اقلیمی نه تنها به نوع الگوریتم مورد استفاده، بلکه به ماهیت فیزیکی و آماری متغیر اقلیمی و ویژگی‌های مکانی ایستگاه وابسته است. یافته‌ها تأکید می‌کند که به کارگیری یک روش واحد برای تمامی متغیرها رویکردی بهینه محسوب نمی‌شود و انتخاب مدل باید به صورت تطبیقی و متغیرمحور انجام گیرد. از منظر کاربردی، نتایج بیان‌گر آن است که در چارچوب داده‌های چندمتغیره اقلیمی، استفاده از مدل‌هایی با توانایی استخراج روابط غیرخطی می‌تواند پایداری بازسازی را بهبود دهد، اما این بهبود زمانی معنادار است که پیچیدگی مدل متناسب با ساختار داده باشد. در غیر این صورت، مدل‌های ساده‌تر و کم‌هزینه‌تر نیز می‌توانند عملکردی قابل قبول ارائه دهند. بنابراین، در کاربردهای عملی نظیر مطالعات اقلیم‌شناسی، تحلیل خشکسالی و مدیریت منابع آب، توجه هم‌زمان به دقت آماری، هزینه محاسباتی و پایداری مدل ضروری است. با این حال، این پژوهش با محدودیت‌هایی همراه است. نخست، تحلیل‌ها بر اساس داده‌های شش ایستگاه در یک حوزه اقلیمی خاص انجام شده و تعمیم نتایج به سایر اقلیم‌ها نیازمند بررسی بیشتر است. دوم، الگوی فقدان داده به صورت کنترل‌شده شبیه‌سازی شد و رفتار مدل‌ها در شرایط فقدان واقعی و غیرتصادفی^۱ می‌تواند متفاوت باشد. سوم، ارزیابی عدم قطعیت بازسازی و بازه‌های اطمینان مقادیر برآوردشده به صورت تفصیلی بررسی نشد که می‌تواند بر کاربردهای تصمیم‌گیری اثرگذار باشد. بر اساس یافته‌های این پژوهش، پیشنهاد می‌شود در تحقیقات آتی، عملکرد مدل‌ها در اقلیم‌های متنوع و با تراکم ایستگاهی متفاوت مورد بررسی قرار گیرد و تأثیر الگوهای مختلف فقدان داده و سطوح بالاتر درصد داده‌های مفقود تحلیل گردد. همچنین، ارزیابی چارچوب‌های پیشرفته‌تر نظیر مدل‌های عمیق یا روش‌های ترکیبی چندمرحله‌ای و انجام نظام‌مند تحلیل کمی عدم قطعیت و حساسیت مدل‌ها به انتخاب پارامترها، گام‌های مهمی در جهت توسعه این حوزه خواهند بود. در مجموع، نتایج این مطالعه بر اهمیت رویکردهای هوشمند و تطبیقی در مدیریت داده‌های ناقص اقلیمی تأکید می‌کند و می‌تواند مبنایی برای بهبود کیفیت پایگاه‌های داده

References

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, 103.455.
doi:10.1016/j.mex.2025.103455
- Amekudzi, L. K., Gyasi-Agyei, Y., Obuobie, E., & Addi, M. (2022). Evaluation of imputation techniques for infilling missing daily rainfall records on river basins in Ghana. *Taylor & Francis*.
- 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. doi:10.1002/mpr.329
- Chivers, B. D., Wallbank, J., Cole, S. J., Sebek, O., Stanley, S., Fry, M., & Leontidis, G. (2020). Imputation of missing sub-hourly precipitation data in a large sensor network: A machine learning approach. *Journal of Hydrology*, 588, 125126. doi:10.1145/2939672.2939785
- Davari, S., Eslamian, S., Jamali, M., & Safavi, H. R. (2025). Application of machine learning algorithms for groundwater level prediction in the Najafabad plain. *Scientific Reports*. doi:10.47176/jwss.24.4.42931
- Diouf, S., Deme, A., El Hadji Deme, P. F., & Diouf, I. (2023). An evaluation of the performance of imputation methods for missing meteorological data in Burkina Faso and Senegal. *Afr. J. Environ. Sci. Technol*, 17, 252-274. doi:10.47176/jwss.24.4.42933
- 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. doi:10.1007/s10668-021-01784-4
- Géron, A., (2022). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media, Inc.
- Ghanavati, R., Salajegheh, A., Pourghasemi, H. R., Khalighi Sigaroodi, S., & Keshtkar, H. (2025). Evaluation of machine learning techniques (SVM, GLM, FDA, RF) in preparing flood susceptibility map of a part of Khuzestan province. *Water and Soil Management and Modeling*, 5(1), 231-246. doi:10.47176/jwss.24.4.42905
- 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. doi: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. doi:10.1061/(ASCE)1084-0699
- Hamzah, F. B., Mohamad Hamzah, F., Mohd Razali, S. F., & El-Shafie, A. (2022). Multiple imputations by chained equations for recovering missing daily streamflow observations: A case study of Langat River basin in Malaysia. *Hydrological Sciences Journal*, 67(1), 137-149.
- Hassanzadeh, E., Zarghami, M. and Hassanzadeh, Y., (2012). Determining the main factors in declining the Urmia Lake level by using system dynamics modeling. *Water Resources Management*, 26(1), pp.129-145.
- 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. doi:10.1016/j.jag.2022.102701
- 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), 29. (In Persian) doi:10.47176/jwss.24.4.42931
- Jamali, M., & Eslamian, S. (2023). Parametric and nonparametric methods for analyzing the trend of extreme events. In *Handbook of Hydroinformatics* (pp. 223-237). Elsevier.
- Jamali, M., & Eslamian, S. (2024). Climate Adaptation and Water-Environment-Energy Nexus. In *Handbook of Climate Change Impacts on River Basin Management* (pp. 225-236). CRC Press.
- Jing, X., Luo, J., Wang, J., Zuo, G., & Wei, N. (2022). A multi-imputation method to deal with hydro-meteorological missing values by integrating chain equations and random forest. *Water Resources Management*, 36(4), 1159-1173.
- Karbasi, M., 2016. Reconstruction of missing data of monthly total sunshine hours using artificial neural networks. *Iranian Journal of Irrigation & Drainage*, 10(5), pp.570-580.
- 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.
- 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 System. Science*, doi:10.5194/hess-23-4323-2019

- Kumar, G. P., & Dwarakish, G. (2025). Comparison of the multiple imputation approaches for imputing rainfall data: A humid tropical river basin case study. *Water Conservation Science and Engineering*, 10(2), 87.
- 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. doi:10.1029/1998WR900018
- Little, R., & Rubin, D. (1987). Multiple imputation for nonresponse in surveys. *Wiley*, 10, 9780470316696.
- Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data*. John Wiley & Sons .
- Liu, J., Jiang, L., Zhang, X., Druce, D., Kittel, C. M., Tøttrup, C., & Bauer-Gottwein, P. (2021). Impacts of water resources management on land water storage in the North China Plain: Insights from multi-mission earth observations. *Journal of Hydrology*, 603, 126933.
- Matinzadeh, M. m., Fattahi, R., Shayan-zadeh, 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://ijwmsei.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. doi:10.13031/2013.23153
- O'Sullivan, B., & Kelly, G. (2024). Infilling of high-dimensional rainfall networks through multiple imputation by chained equations. *International Journal of Climatology*, 44(9), 3075-3091.
- 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. doi:10.1016/j.uclim.2025.102496
- Sahoo, A. and Ghose, D. K. 2022. Imputation of missing precipitation data using KNN, SOM, RF, and FNN. *Soft Computing*. 26: 5919-5936.
- Qaraghuli, K., Murshed, M., Said, M. A. M., Mokhtar, A., & Roustia, I. (2024). Univariate and multivariate imputation methods evaluation for reconstructing climate time series data: A case study of Mosul station-Iraq. *Journal of Agrometeorology*, 26(3), 318-323.
- Tashman, L.J., (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, 16(4), pp.437-450.
- Van Buuren, S. (2000). *Multivariate imputation by chained equations: MICE V1 . User's Manual*. Leiden: TNO .
- Wilks, D.S., 2011. Statistical methods in the atmospheric sciences (Vol. 100). Academic press.
- Willmott, C. J. (1981). ON THE VALIDATION OF MODELS. *Physical Geography*, 2(2), 184-194. doi:10.1080/02723646.1981.10642213