Vol 16, No. 1, Spring 2026

University of Tehran Press

Journal of Water and Irrigation Management Online ISSN: 2382-9931

Homepage: https://jwim.ut.ac.ir/

Reconstruction of missing daily streamflow data using Multiple Imputation by Chained Equations in Kajo river

Fatemeh Ganji Gohari[1] | Hamid Nazaripour[2]✉ | Mohammadreza Poodineh[3] | Mohsen Hamidianpour[4] | Alireza Ghaemi[5] | Reza Teimouri[6]

1. Department of Physical Geography, Faculty of Geography and Environmental Planning, University of Sistan and Baluchestan, Zahedan, Iran. E-mail: fatemeh.ganji77@gmail.com
2. Corresponding Author, Department of Physical Geography, Faculty of Geography and Environmental Planning, University of Sistan and Baluchestan, Zahedan, Iran. E-mail: h.nazaripour@gep.usb.ac.ir
3. Department of Physical Geography, Faculty of Geography and Environmental Planning, University of Sistan and Baluchestan, Zahedan, Iran. E-mail: mr.poodineh@gep.usb.ac.ir
4. Department of Physical Geography, Faculty of Geography and Environmental Planning, University of Sistan and Baluchestan, Zahedan, Iran. E-mail: mhamidianpour@gep.usb.ac.ir
5. Department of Physical Geography, Faculty of Geography and Environmental Planning, University of Sistan and Baluchestan, Zahedan, Iran. E-mail: alireza_ghaemi@pgs.usb.ac.ir
6. Department of Water Engineering, Faculty of Water and Soil Engineering, Gorgan University of Agricultural Sciences and Natural Resources, Gorgan, Iran. E-mail: r.teimourey@hwstre.ir

Article Info

Article type:
Research Article

Article history:
Received 2 Janaury 2026
Received in revised form
6 February 2026
Accepted 10 February 2026
Published online 22 June 2026

Keywords:
Missing Data
Streamflow
Multiple Imputation
MICE
CART

ABSTRACT

Missing values in hydrology studies are a common challenge for hydrologists, especially in statistical analyses that require complete datasets. This research evaluates the performance of the Multiple Imputation by Chained Equations (MICE) method in predicting and reconstructing daily river flow values. The study area is the Kajo River basin in southeastern Iran, and the statistical period covers the hydrological years from 1972-1973 to 2021-2022. To investigate and validate the effectiveness of the MICE approach in managing missing flow data, complete historical daily flow records from the hydrological years 2011–2012 to 2021–2022 were used. Subsequently, the MICE method along with Multiple Linear Regression (MLR) was applied to reconstruct all missing daily flow values. The best-performing estimation methods were evaluated using criteria such as the adjusted coefficient of determination ($Adj\ R^2$), residual standard error (RSE), and mean absolute percentage error (MAPE). The findings indicated that the Classification and Regression Trees (CART) method combined with MLR outperformed other tested methods, achieving the highest $AdjR^2$ value and the lowest RSE and MAPE values. The RSE and MAPE values for the CART-MLR method at the Pirsehrab station are 0.472 and 0.583, respectively, and at the Chandokan station are 0.475 and 0.588, respectively.

Cite this article: Ganji Gohari, F., Nazaripour, H., Poodineh, M., Hamidianpour, M., Ghaemi, A., & Teimouri, R. (2026). Reconstruction of missing daily streamflow data using Multiple Imputation by Chained Equations in Kajo river. *Journal of Water and Irrigation Management*, 16 (1), 141-162.
DOI: https://doi.org/10.22059/jwim.2026.408919.1280

Publisher: University of Tehran Press.
DOI: https://doi.org/10.22059/jwim.2026.408919.1280

انتشارات دانشگاه تهران

مدیریت آب و آبیاری
Homepage: https://jwim.ut.ac.ir/

شاپا الکترونیکی: ۲۳۸۲-۹۹۳۱

بازسازی داده‌های گم‌شده جریان روزانه رودخانه کاجو با انتساب چندگانه معادلات زنجیره‌ای

فاطمه گنجی‌گوهری[۱] | حمید نظری‌پور[۲]✉ | محمدرضا پودینه[۳] | محسن حمیدیان‌پور[۴] | علیرضا قائمی[۵] | رضا تیموری[۶]

۱. گروه جغرافیای طبیعی، دانشکده جغرافیا و برنامه‌ریزی محیطی، دانشگاه سیستان و بلوچستان، زاهدان، ایران. رایانامه: fatemeh.ganji77@gmail.com
۲. نویسنده مسئول، گروه جغرافیای طبیعی، دانشکده جغرافیا و برنامه‌ریزی محیطی، دانشگاه سیستان و بلوچستان، زاهدان، ایران. رایانامه: h.nazaripour@gep.usb.ac.ir
۳. گروه جغرافیای طبیعی، دانشکده جغرافیا و برنامه‌ریزی محیطی، دانشگاه سیستان و بلوچستان، زاهدان، ایران. رایانامه: mr.poodineh@gep.usb.ac.ir
۴. گروه جغرافیای طبیعی، دانشکده جغرافیا و برنامه‌ریزی محیطی، دانشگاه سیستان و بلوچستان، زاهدان، ایران. رایانامه: mhamidianpour@gep.usb.ac.ir
۵. گروه جغرافیای طبیعی، دانشکده جغرافیا و برنامه‌ریزی محیطی، دانشگاه سیستان و بلوچستان، زاهدان، ایران. رایانامه: alireza_ghaemi@pgs.usb.ac.ir
۶. گروه مهندسی آب، دانشکده مهندسی آب و خاک، دانشگاه علوم کشاورزی و منابع طبیعی گرگان، گرگان، ایران. رایانامه: r.teimourey@hwstre.ir

اطلاعات مقاله

نوع مقاله: مقالهٔ پژوهشی

تاریخ دریافت: ۱۴۰۴/۱۰/۱۲
تاریخ بازنگری: ۱۴۰۴/۱۱/۱۷
تاریخ پذیرش: ۱۴۰۴/۱۱/۲۱
تاریخ انتشار: ۱۴۰۵/۰۴/۰۱

کلیدواژه‌ها:
داده گم‌شده
جریان رودخانه
انتساب‌های چندگانه
معادلات زنجیره‌ای
درختان طبقه‌بندی و رگرسیون

چکیده

مقادیر گم‌شده در مطالعات هیدرولوژی یک چالش رایج برای هیدرولوژیست‌ها محسوب می‌شود، به‌ویژه در تحلیل‌های آماری که به مجموعه داده‌های کامل نیاز است. این پژوهش، عملکرد روش انتساب چندگانهٔ جایگزینی مبتنی بر معادلات زنجیره‌ای (MICE) را در پیش‌بینی و بازسازی مقادیر جریان روزانه رودخانه ارزیابی می‌کند. محدوده موردمطالعه پژوهش، حوضه آبریز رودخانه کاجو در جنوب‌شرق ایران و دوره آماری آن در بازه سال‌های هیدرولوژیکی ۱۳۵۲–۱۳۵۱ تا ۱۴۰۱–۱۴۰۰ می‌باشد. برای بررسی و اعتبارسنجی کارایی رویکرد MICE در مدیریت داده‌های گم‌شده جریان، از داده‌های تاریخی روزانه کامل جریان در بازه سال‌های هیدرولوژیکی ۱۳۹۲–۱۳۹۱ تا ۱۴۰۱– ۱۴۰۰ استفاده شده است. در ادامه، روش‌های MICE همراه با رگرسیون خطی چندگانه (MLR) برای بازسازی کُل مقادیر گم‌شده جریان روزانه به‌کار گرفته شده است. برترین روش‌های برآورد با معیارهایی از قبیل ضریب تعیین تعدیل‌شده ($Adj\ R^2$)، خطای استاندارد (RSE) و میانگین درصد خطای مطلق (MAPE) ارزیابی شده است. یافته‌ها نشان داد که روش درختان طبقه‌بندی و رگرسیون (CART) ترکیب‌شده با MLR با بالاترین مقدار $Adj\ R^2$ و کم‌ترین مقادیر RSE و MAPE در مقایسه با سایر روش‌های آزمون شده دارای عملکرد بهتری بوده است. مقادیرRSE و MAPE روش CART-MLR، در ایستگاه پیرسهراب به‌ترتیب۰/۴۷۲ و ۰/۵۸۳ و در ایستگاه چندوکان ۰/۴۷۵ و ۰/۵۸۸ می‌باشد.

استناد: گنجی‌گوهری، فاطمه؛ نظری‌پور، حمید؛ پودینه، محمدرضا؛ حمیدیان‌پور، محسن؛ قائمی، علیرضا و تیموری، رضا (۱۴۰۵). بازسازی داده‌های گم‌شده جریان روزانه رودخانه کاجو با انتساب چندگانه معادلات زنجیره‌ای. *نشریه مدیریت آب و آبیاری*، ۱۶ (۱)، ۱۴۱–۱۶۲.
DOI: https://doi.org/10.22059/jwim.2026.408919.1280

ناشر: مؤسسه انتشارات دانشگاه تهران.

۱. مقدمه

یکی از چالش‌های متداول در تحقیقات هیدرولوژی، وجود داده‌های گم‌شده در مجموعه داده‌هاست. با وجود معرفی رویکردهای متنوع بازسازی داده‌های گم‌شده طی سال‌های اخیر، مسئله مقادیر گم‌شده که تحلیل‌های هیدرولوژیکی را محدود می‌سازد در نتیجه وقوع بلایای طبیعی یا عملکرد نامناسب تجهیزات (Mwale *et al.,* 2012)، همچنان پابرجاست (Mispan *et al.,* 2015; Tencaliec *et al.,* 2015; Hamzah *et al.,* 2020). پیچیدگی‌های فنی، شرایط نامساعد جوی، خرابی تجهیزات یا خطاهای ابزاری در فرایند جمع‌آوری داده، اشتباه کاربر در هنگام ورود داده، آسیب‌دیدگی داده‌ها در نتیجه عملکرد نامناسب دستگاه‌های ذخیره‌سازی، هم‌چنین کاهش بودجه، مشکلاتی در ساختاردهی و سازمان‌دهی بلندمدت داده‌های هیدرومتری ایجاد کرده و در مواردی منجر به ایجاد شکاف در مجموعه داده‌ها می‌شوند (Johnston, 1999; Gao, 2017; Tencaliec, 2017; Gires *et al.,* 2021). داده‌های مفقوده به‌ویژه در حوضه‌های آبریز دورافتاده مشاهده می‌شود که در آن‌ها خرابی تجهیزات تنها پس از تأخیرهای قابل‌توجهی که متعاقب رویدادهای حدی رخ می‌دهد، تعمیر می‌گردد؛ امری که می‌تواند برای تحلیل‌های فرکانسی هیدرولوژیک حیاتی باشد (Ahn, 2021). پیامدهای استفاده از چنین داده‌هایی، عدم قطعیت و کاهش کارایی سیستم‌های مدیریت منابع آب می‌باشد (Adeloye, 1996).

رویکرد پذیرفته‌شده در مدل‌سازی هیدرولوژیک، حذف مشاهداتی است که در هر بازه زمانی فاقد مقادیر متغیرها هستند، حتی اگر تنها یکی از متغیرهای مستقل فاقد داده باشد (Gill *et al.,* 2007). معمولاً داده‌های ناقص نشانه‌گذاری‌شده و از فرایند ساخت مدل و هم‌چنین آزمون و اعتبارسنجی متعاقب آن حذف می‌گردند. با این‌حال، این روش نشان‌دهنده فقدان رویکرد مناسب در مواجهه با داده‌های گم‌شده است که می‌تواند منجر به ایجاد بایاس و یا ازدست‌رفتن اطلاعات ارزشمند گردد. این امر به‌نوبه خود ممکن است بر تفسیر داده‌ها، کارایی تحلیلی و یافته‌های علمی تأثیرگذار باشد (Zhao & Long, 2016; Semiromi & Koch, 2019; Nor *et al.,* 2020). حتی شکاف‌های داده‌ای بسیار کوچک، ممکن است محاسبه قابل‌توجه آماره‌های ضروری و شاخص‌های هیدرولوژیکی را منتفی کنندو از این‌رو، تجزیه و تحلیل و توضیح تغییرپذیری جریان گذشته را محدود می‌کند (Harvey *et al.,* 2012). بنابراین، بازسازی و پردازش داده‌های گم‌شده می‌باید در گام نخست فرایند آماده‌سازی داده‌ها انجام پذیرد، جایی که رویکرد مورداستفاده، تحت‌تأثیر الگو و سازوکار فقدان داده قرار می‌گیرد (Plaia & Bondì, 2006; Ahmat Zainuri, *et al.,* 2015; Kamaruzaman *et al.,* 2017).

داده‌های گم‌شده دارای سه نوع می‌باشند؛ ۱- داده گم‌شده کاملاً تصادفی (MCAR)، ۲- داده گم‌شده تصادفی (MAR)، ۳- داده گم‌شده غیرتصادفی (MNAR) (Little & Rubin, 2002). سازوکار داده گم‌شده به‌عنوان گمشدگی کاملاً تصادفی شناخته می‌شود که کاملاً مستقل از مقادیر هر یک از متغیرهای موجود در مجموعه‌داده‌هاست، خواه این مقادیر مفقوده باشند یا مشاهده‌شده. بنابراین این نوع از گمشدگی در نتیجه یک اتفاق کاملاً تصادفی رخ می‌دهد. در مقابل، گمشدگی تصادفی که می‌توان به‌عنوان ریشه گمشدگی داده‌ها نیز توصیف کرد با مقادیر گم‌شده مرتبط نمی‌باشد، اما ممکن است با مقادیر مشاهده سایر متغیرها همبستگی داشته باشد. در نهایت، گمشدگی غیرتصادفی، به‌طور تصادفی نمی‌باشند و بنابراین در دسته MCAR و MAR قرار نمی‌گیرند. این طبقه‌بندی در انتخاب روش صحیح بازسازی داده‌های مفقوده، حیاتی است. بازسازی داده‌های جریان رودخانه با فرض گمشدگی تصادفی توسط Gill و همکاران (۲۰۰۷) انجام شده است. با این‌حال، با استناد به تعریف Little & Rubin (۲۰۰۲)، داده‌های گم‌شده در مطالعات جریان رودخانه می‌توانند به‌عنوان کاملاً تصادفی در نظر گرفته شوند، چرا که فقدان داده در جریان‌سنجی یک منطقه، ممکن است تحت تأثیر داده‌های همان منطقه یا سایر مناطق قرار نگیرد.

در سال‌های اخیر، علاقه فزاینده‌ای به بازسازی داده‌های گم‌شده جریان رودخانه با استفاده از رویکردهای آماری

متعدد پدیدار شده است (Regonda *et al.,* 2013). برای مواجهه با مسئله مقادیر گم‌شده، روش‌های متعدد برآورد داده‌ها در ادبیات موضوع پیشنهاد و به‌طور گسترده موردبحث قرار گرفته‌اند. این روش‌ها از ابتدایی‌ترین تکنیک‌های آماری سنتی (مانند جایگزینی مقادیر مفقوده با میانگین، میانه یا داده‌های ایستگاه‌های دیگر) تا روش‌های محاسباتی پیشرفته را در بر می‌گیرند. در میان رویکردهای آماری طراحی‌شده برای بازسازی داده‌های مفقوده، روش انتساب چندگانه (MI) را می‌توان در شرایط متنوعی با استفاده از بسته‌های نرم‌افزاری موجود اجرا نمود. این روش به پژوهش‌گر امکان می‌دهد تا تحلیل‌های استاندارد مبتنی بر داده‌های کامل را به‌طور مستقیم بر روی مجموعه‌داده‌های بازسازی شده اعمال کرد. در روش MI داده‌های گم‌شده n بار با استفاده از یک مدل آماری پُر می‌شوند و در نتیجه n مجموعه داده کامل به‌دست می‌آید که برای تحلیل قابل استفاده می‌باشند. ایده اصلی این است که هر داده گم‌شده با دو یا چند مقدار محتمل جایگزین شود که نشان‌دهنده توزیعی از احتمالات باشند. یک تکنیک شناخته‌شده در اجرای MI، مدل‌سازی رگرسیون ترتیبی است که به‌عنوان روش انتساب چندگانه با معادلات زنجیره‌ای (MICE) شناخته می‌شود. صرف‌نظر از محدودیت، این تکنیک با توجه به انعطاف‌پذیری و سادگی نسبی پیاده‌سازی آن، به‌طور گسترده مورداستفاده قرار می‌گیرد (Su *et al.,* 2011; van Buuren & Groothuis-Oudshoorn, 2011).

علاوه بر روش MICE، مطالعات متنوع دیگری نیز در مورد روش‌های پیش‌بینی و بازسازی داده‌های جریان مفقوده انجام شده است. به‌عنوان نمونه، از مدل‌های جنگل تصادفی (RF) و ماشین بردار پشتبان (SVM)، برای پیش‌بینی جریان ماهانه رودخانه مارون استفاده و دقت آن‌ها موردارزیابی قرار گرفته است (Nekoeeyan *et al*., 2022). مدل Patch-TST از مجموعه معماری ترنسفورمر و LSTM از مجموعه تکنیک‌های یادگیری عمیق در پیش‌بینی جریان روزانه رودخانه سفیدرود استفاده شده است. نتایج نشان داده است که مدل Patch-TST عملکرد برتری نسبت به مدل LSTM داشته است (Feizi & Sattari, 2026). الگوریتم امید ریاضی- بیشینه‌سازی (EM)، شبکه عصبی مصنوعی (ANN) و MICE برای بازسازی داده‌های گم‌شده بارش در ایستگاه کوانتان (مالزی) استفاده شده است. یافته‌های این پژوهش نشان داده است که روش ANN به‌عنوان گزینه برتر شناخته شده است (Norazizi & Deni, 2019). مطالعه مشابهی نیز توسط Zvarevashe و همکاران (۲۰۱۹) با استفاده از روش MICE برای بازسازی داده‌های گم‌شده بارش انجام شده است. روش MICE، توزیع نرمال داده‌ها را مفروض نمی‌گیرد و گمشدگی داده‌ها را به‌عنوان تصادفی در نظر می‌گیرد. روش جدید انتساب واحد به‌نام روش اثر وابسته به مکان برای مقادیر گم‌شده در مجموعه داده‌های آلودگی محیط زیست در شهر پالرمو (ایتالیا) مورداستفاده قرار گرفته و عملکرد آن با سایر روش‌های انتساب مرسوم مقایسه شده است. نتایج نشان داده است که همه شاخص‌های عملکرد، روش پیشنهادی را به‌عنوان بهترین روش در بین روش‌های مقایسه‌شده، به‌طور مستقل براساس طول شکاف و تعداد ایستگاه‌های دارای داده‌های گم‌شده، ارزیابی کرده است (Plaia & Bondì, 2006). مطالعات متعددی نشان داده‌اند که استفاده از روش MICE منجر به عملکرد پیش‌بینی بهتری در مدل‌های طبقه‌بندی یا پیش‌بینی می‌شود (Donders *et al.,* 2006; Schmitt *et al.,* 2015; Chhabra *et al.,* 2017). روش MICE برای مجموعه داده‌های بالینی نسبت به روش انتساب واحد ترجیح داده شده است. زیرا در انتساب واحد ، فقط از یک تخمین استفاده می‌شود. درحالی‌که در انتساب چندگانه، از تخمین‌های مختلفی استفاده می‌شود که نشان‌دهنده عدم قطعیت در تخمین این توزیع است. در این مطالعه یک شبیه‌سازی ساده نشان داده است که روش MICE خطای استاندارد و فواصل اطمینان را به‌خوبی برآورد می‌کند (Donders *et al.*, 2006). در مطالعه دیگر، گزارش شده است که الگوی MICE در مقایسه با سایر روش‌های انتساب که از فرض گمشدگی کاملاً تصادفی بهره می‌برند، عملکرد آن به اندازه مجموعه داده وابسته است (Schmitt *et al.,* 2015). قدرت روش MICE در به‌دست‌آوردن خطاهای استاندارد کوچک‌تر و فواصل

اطمینان باریک‌تر نهفته است که در آن‌ها می‌توان مقادیر پیش‌بینی‌شده دقیق‌تری را به‌دست آورد، بنابراین سوگیری و ناکارآمدی را به میزان قابل‌توجهی به حداقل می‌رساند. علاوه بر این، در این مطالعه پیشنهاد شده است که ترکیب روش‌های MICE با یادگیری ماشین و الگوریتم‌های ژنتیک می‌تواند سوگیری و ناکارآمدی را بیش‌تر محدود کند (Chhabra *et al.*, 2017). اگرچه پژوهش‌های قابل‌توجهی در زمینه بازسازی مقادیر گم‌شده با استفاده از روش MICE در شرایط آزمایشی مختلف انجام شده است، تنها تعداد معدودی از مطالعات به بازسازی داده‌های گم‌شده جریان رودخانه با استفاده از این روش پرداخته‌اند. روش‌های دیگر مورداستفاده برای انتساب داده‌های گم‌شده جریان رودخانه شامل مدل رگرسیون بیزی سلسله مراتبی است که از آن برای بازسازی میانگین جریان تابستانه در پنج ایستگاه اندازه‌گیری در حوضه رودخانه دلاور استفاده شده است (Devineni *et al.,* 2013). هم‌چنین، در برخی مطالعات استفاده از روش‌های درخت‌های چندگانه طبقه‌بندی و رگرسیون (CART) یا جنگل تصادفی برای انتساب داده‌های گم‌شده جریان رودخانه توصیه شده است (Vezza *et al.,* 2010; Erdal & Karakurt, 2013; Karakurt *et al.,* 2013; Tyralis *et al.,* 2019) و یافته‌ها نشان داده است که مدل CART در مورد واریانس توضیح داده شده، عملکرد بهتری نسبت به مدل‌های مشتق‌شده از سایر روش‌های طبقه‌بندی دارد.

رویکرد بسته MICE برای ارزیابی داده‌های گم‌شده از سری‌زمانی تک متغیره با استفاده از انتساب چندگانه مورداستفاده قرار گرفته است. با این وجود، برای تخمین جریان رودخانه، مقایسه بین رویکردهای MICE هنوز انجام نشده است، به‌ویژه استفاده از تطبیق میانگین پیش‌بینی‌کننده (PMM)، تخمین رگرسیون تصادفی (SRI) و رگرسیون خطی چندگانه با تخمین بوت‌استرپ (BOOT) برای بازسازی داده‌های جریان رودخانه گم‌شده. در پژوهش‌های هیدرولوژیک، ایجاد مدل‌هایی که چندین متغیر را در بر می‌گیرند چالش‌برانگیز است، زیرا روابط بین متغیرها ممکن است تعاملی و غیرخطی باشند و تشخیص این پیچیدگی‌ها می‌تواند کاری دشوار و بدون تضمین موفقیت باشد. علاوه بر این، بسیاری از متغیرها دارای توزیع‌هایی هستند که به‌دشواری با استفاده از مدل‌های پارامتریک استاندارد قابل شناسایی هستند. در نتیجه، هدف اول این مطالعه بررسی دقت بسته نرم‌افزاری MICE (van Buuren & Groothuis-Oudshoorn, 2011) با استفاده از مدل‌های شرطی مانند PMM، SRI، CART، BOOT و انتساب رگرسیون خطی بیزی برای تخمین رکوردهای جریان گم‌شده می‌باشد. هدف دوم، ارزیابی عملکرد روش‌های انتساب در ترکیب با مدل رگرسیون خطی چندگانه (MLR) در پیش‌بینی مقادیر روزانه جریان رودخانه ارزیابی خواهد بود. انتساب با روش MICE هم‌چنین می‌تواند برای جایگزینی داده‌های گم‌شده جریان رودخانه بدون نیاز به اطلاعات از ایستگاه‌های پایش مجاور مورداستفاده قرار گیرد. انتظار می‌رود که یافته‌های این مطالعه منجر به شناسایی بهترین و دقیق‌ترین روش‌ها برای انتساب داده‌ها کمک کند، که امکان بازسازی مجموعه داده‌های کامل روزانه جریان رودخانه را فراهم می‌سازد.

۲. روش‌شناسی پژوهش

۲. ۱. منطقه موردمطالعه

محدوده موردمطالعه این پژوهش، حوضه آبریز رودخانه کاجو (شکل ۱) را در برمی‌گیرد. حوضه آبزیز رودخانه کاجو با مساحت ۶۱۷۸ کیلومتر مربع در بخش غربی حوضه آبریز بزرگ باهوکلات و یکی از مهم‌ترین سرشاخه‌های رودخانه باهوکلات در حوضه آبریز بلوچستان جنوبی به‌شمار می‌رود. این رودخانه از ارتفاعات رشته کوه مَکُران در جنوب‌غربی ایرانشهر سرچشمه و در جلگه دشتیاری به رودخانه باهوکلات می‌پیوند. طول آبراهه اصلی این حوضه آبریز ۳۰۲ کیلومتر، زمان تمرکز آن ۴۳/۵ ساعت و ضریب گراویلیوس آن ۲/۸۶ می‌باشد. رودخانه کاجو یکی از بسترهای سیلابی مهم در

جنوب‌شرق ایران شناخته می‌شود و سد مخزنی زیردان بر روی آبراهه اصلی آن احداث شده است. تفاوت زیاد مرتفع‌ترین (۲۰۶۲ متر) و پست‌ترین (۱۶ متر) نقطه حوضه آبریز، شیب زیاد (متوسط ۱۶ درصد)، خشکی محیط جغرافیایی (فقدان پوشش گیاهی) و کافی نبودن اقدامات کنترلی، منجر به رخداد سیلاب‌های سهمگین می‌گردد که خسارت‌های قابل‌توجهی را به‌بار می‌آورد. حوضه آبریز رودخانه کاجو از نظر آب و هواشناختی تحت تأثیر دو رژیم بارشی متفاوت قرار داد. رژیم موسمی با جهت جریان جنوب‌شرقی از اواخر بهار تا اواخر تابستان و رژیم مدیترانه‌ای که با جهت جریان جنوب‌غربی از میانه پاییز تا میانه بهار حاکم است (Aryanmanesh *et al.*, 2024).

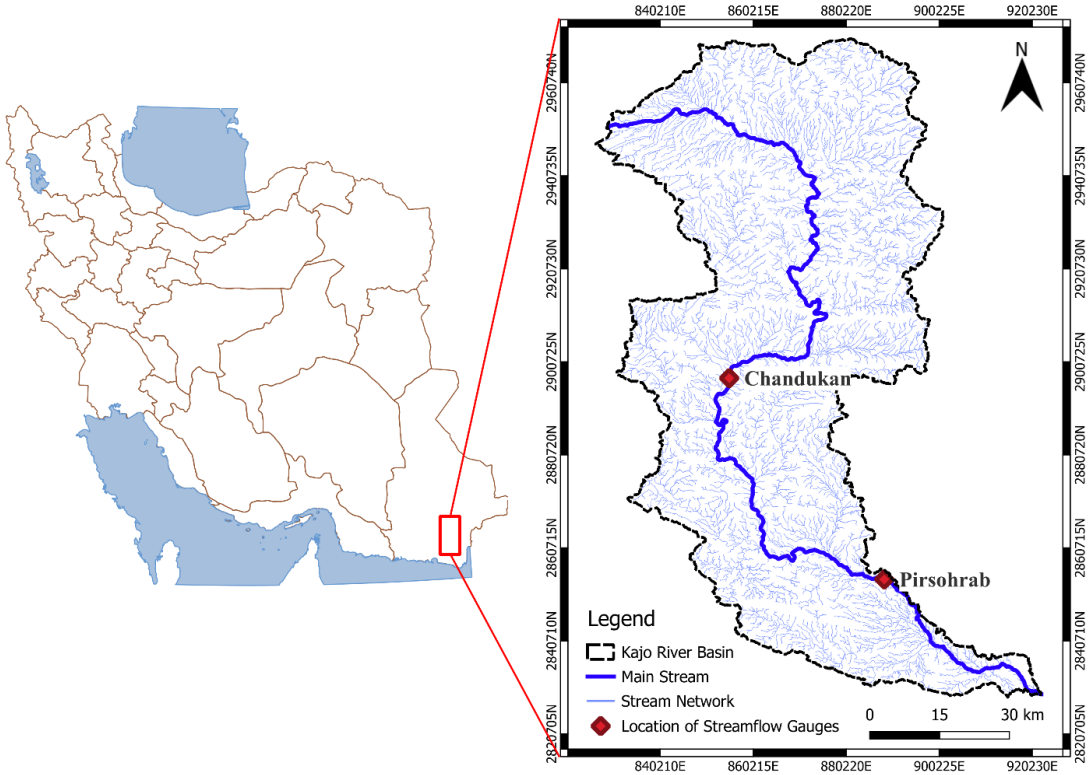

Figure 1. Map of the Kajo River basin and location of streamflow gauges

۲. ۲. داده‌ها

داده‌های این پژوهش شامل جریان روزانه رودخانه در دو ایستگاه هیدرومتری واقع بر روی رودخانه کاجو شامل ایستگاه پیرسهراب در بخش پایاب و ایستگاه چندوکان در بخش سرآب (بالادست سد زیردان) می‌باشد. ایستگاه پیرسهراب در موقعیت طول جغرافیایی ۶۰/۸۷ و عرض جغرافیایی ۲۵/۷۵ و ایستگاه چندوکان در موقعیت طول جغرافیایی ۶۰/۵۵ و عرض جغرافیایی ۲۶/۱۵ قرار دارد. دوره آماری مورداستفاده هر دو ایستگاه از سال آبی ۱۳۵۱/۱۳۵۲ تا ۱۴۰۱/ ۱۴۰۰ می‌باشد. میانگین درصد گمشدگی داده‌های جریان روزانه در ایستگاه پیرسهراب ۴۳/۶ و در ایستگاه چندوکان ۳۷/۸ درصد می‌باشد که طبق نظر (Widaman, 2006) در محدوده درصد گمشدگی بالا (۵۰–۲۵ درصد) قرار دارند (جدول ۱). طبق آماره‌های جریان روزانه، ضریب تغییرپذیری جریان روزانه در هر دو ایستگاه به‌شدت بالا بوده و توزیع آن‌ها بسیار چوله به راست بوده که نشان‌دهنده وقوع سیلاب‌های شدید می‌باشد. ناپیوستگی ذاتی در رژیم جریان رودخانه کاجو، یکی

از کلیدی‌ترین دلایل بالا بودن ضریب تغییرات جریان می‌باشد. میزان چولگی و کشیدگی در ایستگاه چندوکان به نسبت شدیدتر بوده که حاکی از نوسانات شدیدتر و داده‌های پرت بسیار زیاد در این ایستگاه می‌باشد (جدول ۱). توزیع جریان روزانه گسترده‌تر و با مقادیر حداکثری بیش‌تر نسبت به ایستگاه چندوکان مشخص می‌گردد. کشیدگی مثبت بالا نیز بیانگر مقادیر زیاد جریان‌های سیلابی در هر دو ایستگاه آب‌سنجی می‌باشد (شکل ۲). تفاوت معنی‌داری بین درصد گمشدگی در ماه‌های سال وجود ندارد. ماه‌های فصل زمستان نسبت به سایر ماه‌ها، میزان گمشدگی پایین‌تری دارند (شکل ۳). فراوانی نسبی گروه‌های مشخص از طول گمشدگی آماری (روز) نیز در شکل (۳) نمایش داده شده است. گمشدگی آماری با طول ۱۴، ۲۱، ۳۰، ۶۰ و ۹۰ روزه در هر دو ایستگاه آب‌سنجی، سهم بالاتری از کل گمشدگی را به خود اختصاص داده‌اند. درک این ویژگی‌ها در سازوکار بازسازی داده‌های گم‌شده بسیار حائز اهمیت می‌باشد. شاخصه‌های درصد گمشدگی، طول گمشدگی، رژیم جریان و اندازه مشاهدات در انتخاب روش‌های بازسازی و میزان انتظار موفقیت آن‌ها بسیار تأثیرگذار می‌باشد.

Table 1. Descriptive statistics of daily streamflow ($m^3.s^{-1}$) in Kajo River basin

Station	Missingness	Mean	Median	S	Maximum	Kurtosis	Skewness	CV (%)
Pirsohrab	43.6 %	7.74	1.2	35.15	1580.76	625.68	20.41	454.1
Chandokan	37.8 %	3.73	0.5	21.23	1188	1081.58	26.95	569.8

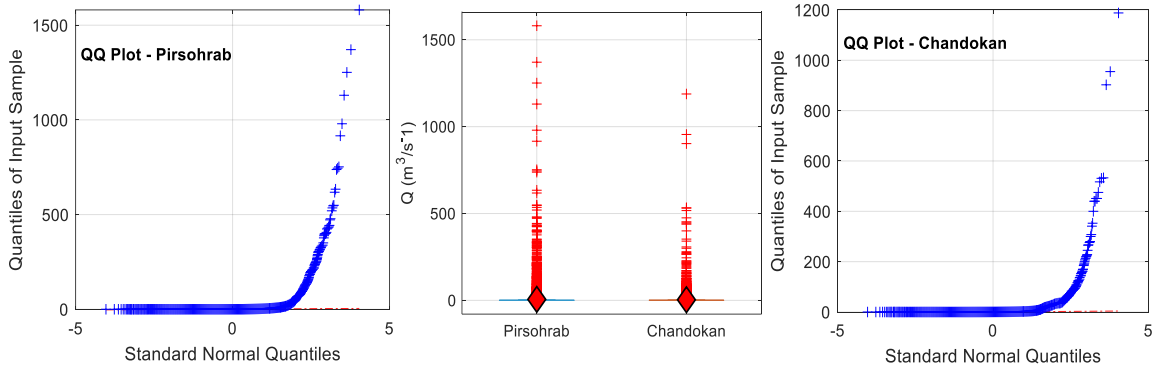

Figure 2. Distribution (Boxplot) and Normality (QQ Plot) Analysis of Daily Flow Data in the Kajo River Basin

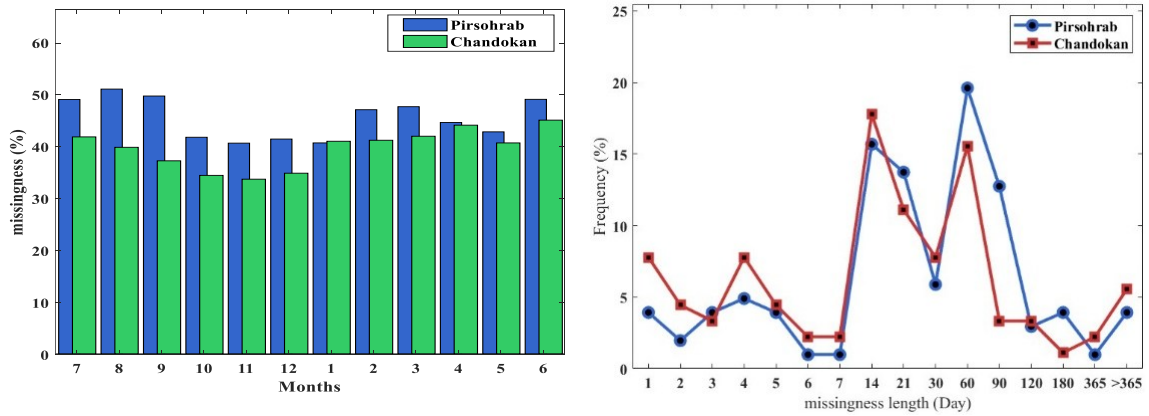

Figure 3. Monthly distribution of missing data (left) and relative frequency of gap lengths (right) in streamflow gauges over the Kajo River basin

۲. ۳. روش‌شناسی

این بخش به دو زیربخش اصلی تقسیم شده است. در زیربخش نخست، روش‌های برآورد داده‌های گم‌شده موردبررسی قرار گرفته و در زیربخش دوم، ارزیابی عملکرد روش‌ها موردبررسی قرار گرفته است. روش مورداستفاده در این مطالعه، یک روش اعتبارسنجی متقاطع برای داده‌های سال‌های ۱۳۹۰/۱۳۹۱ تا ۱۴۰۰/۱۴۰۱ بود تا توانایی روش‌های تکمیل‌داده بررسی شود. دوره ۱۳۹۰/۱۳۹۱ تا ۱۴۰۰/۱۴۰۱ به‌عنوان دوره پایه انتخاب شد، زیرا داده‌های کامل برای این بازه زمانی در دسترس بود. مراحل بازسازی داده‌های گم‌شده جریان روزانه در درون یک سری زمانی کامل در شکل (۴) به تصویر کشیده شده است.

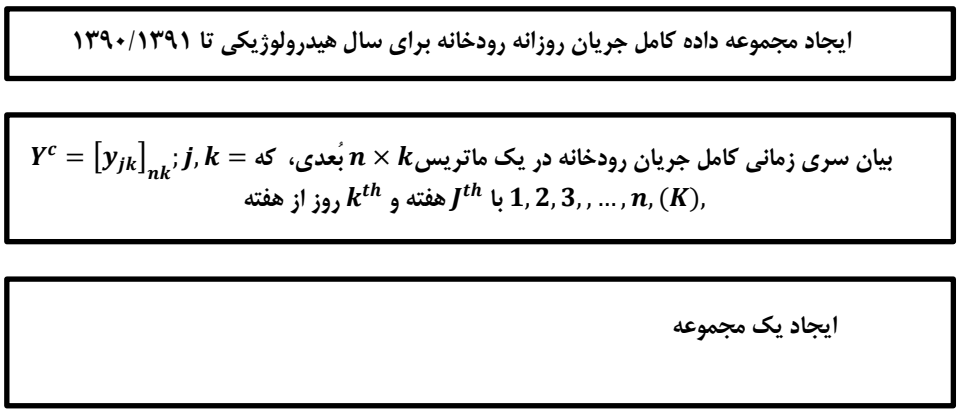

Figure 4. The procedure for introducing the missing data into the complete time series

در ابتدا، تمام مقادیر گم‌شده با استفاده از روش‌های MICE و با جایگزینی از مقادیر مشاهده‌شده، پُر می‌شوند، همان‌گونه که توسط White & Wood (۲۰۱۱) توضیح داده شده است. متغیر اولیه با مقادیر گمشدگی، مثلاً x_1، بر روی تمام متغیرهای دیگر $x_2, x_3, \dots x_k$ رگرسیون می‌گردد، اما تنها برای مواردی که مقدار x_1 در آن‌ها مشاهده شده است. مقادیر گم‌شده در x_1 با مقادیر شبیه‌سازی‌شده از توزیع پیش‌بینی پسین x_1 جایگزین می‌شوند. متغیر بَعدی با مقادیر گم‌شده، مثلاً x_2 روی همه متغیرهای دیگر $x_1, x_3, \dots x_k$ رگرسیون می‌گردد، با این محدودیت که تنها برای مواردی که که مقدار x_2 در آن‌ها مشاهده شده و در عین‌حال از مقادیر بازسازی‌شده x_1 استفاده می‌گردد. مقادیر گم‌شده در x_2 نیز با مقادیر شبیه‌سازی‌شده از توزیع پیش‌بینی پسین x_2 جایگزین می‌شوند. این فرایند به‌نوبت برای هر متغیر دارای مقادیر گم‌شده تکرار می‌شود که به آن یک چرخه اطلاق می‌گردد. برای تثبیت نتایج، این رویه معمولاً برای چندین چرخه (مثلاً

۱۰ یا ۲۰ بار) تکرار می‌شود تا یک مجموعه داده تکمیل‌شده، تولید و کُل فرایند m بار تکرار می‌گردد تا m مجموعه داده تکمیل‌شده ایجاد گردد. سپس، ضریب تعیین تعدیل‌شده ($Adj\ R^2$)، خطای استاندارد باقی‌مانده‌ها (RSE) و میانگین درصد خطای مطلق (MAPE) برای هر یک از پنج مقدار پیش‌بینی محاسبه می‌گردد. در نهایت، روش‌های MICE در ترکیب با رگرسیون خطی چندگانه (MLR) برای بازسازی جریان روزانه رودخانه در حوضه آبریز رودخانه کاجو در دوره هیدرولوژیکی ۱۳۵۱/۱۳۵۲ تا ۱۳۸۹/۱۳۹۰ مورداستفاده قرار می‌گیرد.

۳. ۲. ۱. روش‌های انتساب (جایگزینی)

روش MI، یک رویکرد نوین در مواجهه با مسائل داده‌های گم‌شده است. روش MI هر مقدار گم‌شده را با چندین جواب معتبر و ممکن جایگزین می‌کند. با کمک روش‌های تکمیل داده‌ها، مجموعه داده ناقص به یک مجموعه داده کامل تبدیل می‌شود که سپس می‌توان با هر روش تحلیل استانداردی آن را بررسی نمود (van Buuren, 2007). در مقایسه با روش انتساب واحد، این روش عدم قطعیت موجود در تخمین مقادیر گم‌شده را نیز در نظر می‌گیرد (Hamzah *et al*., 2021). این روش چندین مجموعه‌داده ایجاد می‌کند که از میان آن‌ها می‌توان پارامترهای موردنظر را برآورد کرد (Chhabra *et al*., 2017). در مقایسه با روش انتساب واحد، برآورد دقیق‌تر و کم‌خطاتری از واریانس را ارائه می‌دهد.

در این پژوهش، پنج روش MICE شامل PMM، SRI، BLR، CART و BOOT، به‌عنوان مدل‌های شرطی برای جایگزینی داده‌ها در برآورد مقادیر گم‌شده جریان مقایسه شده‌اند. شکل (۵) مراحل اصلی به‌کاررفته در روش MICE را نشان می‌دهد که توسط van Buuren & Groothuis-Oudshoorn (۲۰۱۱) پیشنهاد شده است. مزیت روش MICE این است که نتایج آن در طی تعداد نسبتاً کمی از تکرارها محاسبه می‌شود. معمولاً پنج تکرار برای دست‌یابی به نتایج پایدار کافی است (Müller *et al*., 1997).

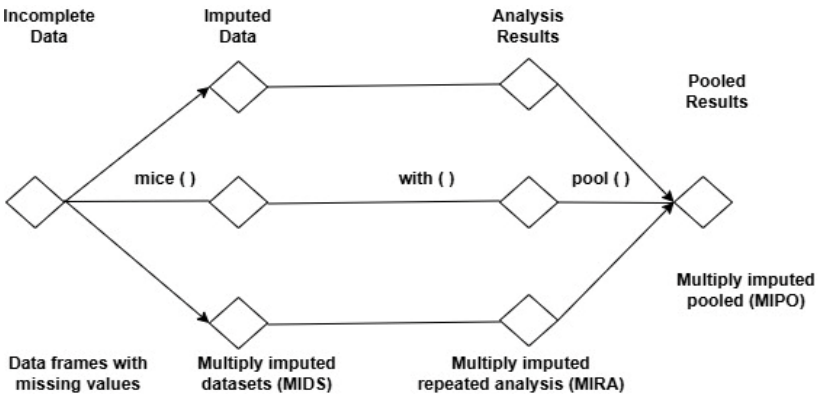

Figure 5. Flowchart of the MICE Algorithm

روش بازسازی با ایجاد یک مدل پیش‌بینی برای متغیر هدفِ دارای مقادیر گم‌شده، بر پایهٔ همهٔ متغیرهای دیگر، انجام شده است. متغیر پاسخ، متغیری است که عملیات تکمیل داده بر روی آن انجام می‌شود و سایر متغیرهای مرتبط، متغیرهای مستقل هستند. معادله (۱) روابط رگرسیونی را نشان می‌دهد:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$ رابطه ۱)

فرض که هر یک از k متغیر مستقل، $x_1, x_2, \dots x_k$، دارای n سطح باشد. در این‌صورت، x_{ij} سطح i^{th} از متغیر مستقل j^{th} را نشان می‌دهد و متغیرهای وابسته، $y_1, y_2, \dots y_k$، نیز دارای n سطح باشند. بنابراین، فرض بر این است که n–تایی مشاهده شده همگی از یک مدل پیروی می‌کنند که به‌صورت معادلات (۲) تا (۵) زیر بیان شده‌اند:

رابطه ۲) $$y_1 = b_0 + b_1 x_{11} + b_2 x_{12} + \cdots + b_k x_{1k} + e_1$$

رابطه ۳) $$y_2 = b_0 + b_1 x_{21} + b_2 x_{22} + \cdots + b_k x_{2k} + e_2$$

رابطه ۴) $$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_k x_{ik} + e_i$$

رابطه ۵) $$y_n = b_0 + b_1 x_{n1} + b_2 x_{n2} + \cdots + b_k x_{nk} + e_n$$

معادله (۱) در قالب معادله (۶) بازنویسی می‌گردد:

رابطه ۶) $$y = X\beta + \varepsilon$$

که در این روابط، X یک ماتریس با ابعاد $(n \times k)$ از n مشاهده روی k متغیر مستقل $x_1, x_2, \dots x_k$، y یک بردار $(n \times 1)$ از n مشاهده مربوط به متغیر وابسته مطالعه، β یک بردار $(k \times 1)$ از ضرایب رگرسیون و ε یک بردار $(n \times 1)$ از خطاها می‌باشد.

با استفاده از نمادگذرای ماتریسی، این n معادله را می‌توان به‌صورت معادله (۷) زیر نوشت:

رابطه ۷)
$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

ستون نخست در ماتریس X مربوط به β_0 و ضریب رگرسیون به‌صورت معادله (۸) بیان می‌گردد:

رابطه ۸) $$\beta = (X'X)^{-1}X'_y$$

که در آن، X' ماتریس ترانهاده از X است.

۳. ۲. ۱. ۱. تطبیق میانگین پیش‌بینی‌کننده

تطبیق میانگین پیش‌بینی کننده (PMM)، یک روش جذاب و پُرکاربرد برای جایگزینی مقادیر گم‌شده در متغیرهای کمّی است (Chhabra *et al.,* 2017). این رویکرد، بر خلاف بسیاری از رویکردهای تکمیل داده، از رگرسیون خطی برای تولید مقادیر جایگزین شده استفاده نمی‌کند. در عوض، یک معیار برای تطبیق موارد دارای داده‌های گم‌شده با موارد مشابه در داده‌های موجود، شناسایی می‌شود. یک فاصله پیش‌بینی‌شده δ_{hj} محاسبه می‌گردد که به‌عنوان معیاری از کیفیت انطباق تعریف می‌گردد. برای تمامی مقادیر j، h مشاهده‌ای که کم‌ترین مقدار $|\delta_{hj}|$ را دارد، براساس معادله (۹) انتخاب می‌گردد:

رابطه ۹) $$\delta_{hj} = \alpha^{mis} z_j - \alpha^{obs} z_h$$

در این حالت، فرض می‌شود که h مشاهداتی را نشان می‌دهد که مقدار x در آن‌ها مشاهده‌شده و j مشاهداتی را نشان می‌دهد که مقدار x در آن‌ها گم‌شده است. برای تمامی h، پیش‌بین خطی $\alpha^{obs} z_h$ و برای تمامی زها پیش‌بین خطی $\alpha^{mis} z_j$ محاسبه می‌گردد. مقادیر مشاهده‌شده اطراف مقدار پیش‌بینی‌شده خطی، به‌عنوان مجموعه اهداگر انتخاب

می‌گردند. معمولاً مجموعه اهداگر به گونه‌ای تعیین می‌شود که شامل k اهداگر کاندید باشد که به‌صورت تصادفی انتخاب می‌شوند.

پرسش اصلی در روش PMM که باید حل شود این است که در هر مجموعه تطبیق، چند مورد k باید وجود داشته باشد. برای تعیین مجموعه اهداگر، سه روش گسترده وجود دارد. نخستین روش، استفاده از تعداد ثابتی از اهداگرها (k) است، در برخی نرم‌افزارها مقدار k=۵ در نظر گرفته می‌شود. به‌طور کلی، هر مورد فردی با داده ناقص روی متغیر x، با پنج مورد کامل که نزدیک‌ترین مقادیر پیش‌بینی شده را دارند، جفت می‌شود. یکی از آن پنج مورد کامل، به‌صورت تصادفی انتخاب می‌گردد و مقدار x آن به مورد دارای داده گم‌شده اختصاص داده می‌شود. روش دوم، تعریف یک حداکثر فاصله δ_{max} است، به‌طوری‌که هر مشاهده h که شرط $|\delta_{hj}| < \delta_{max}$ را برآورده کند، در مجموعه اهداگر برای مشاهده j قرار می‌گیرد، این روش با عنوان همسان‌سازی کالیپر شناخته می‌شود. روش سوم، استفاده از $k = n_h$ است، یعنی تعداد تمام مشاهداتی که مقدار x در آن‌ها مشاهده شده است، با این قابلیت که مقدار مشاهده‌ای که کم‌ترین فاصله d_{hj} را دارد، انتخاب شود.

۳. ۲. ۱. ۲. تخمین رگرسیون تصادفی

روش SRI، مشابه روش انتساب (جایگزینی) رگرسیونی است. در این روش، مقادیر گم‌شده با استفاده از رگرسیون بر سایر متغیرهای مرتبط در همان مجموعه داده، به‌همراه یک ارزش باقیمانده تصادفی، برآورد می‌شوند (Jamil, 2012). به بیان دیگر، روش SRI، مستلزم واردکردن خطای تصادفی به پیش‌بینی رگرسیونی است که از طریق جایگزینی رگرسیونی به‌دست آمده است. معادله (۱۰) مقدار برآوردشده برای داده مفقود را نشان می‌دهد:

$$\dot{y} = \hat{\beta}_0 + X_{mis}\hat{\beta}_1 + \dot{\epsilon} \quad \text{رابطه ۱۰)}$$

که در آن، $\dot{\epsilon}$ به‌صورت تصادفی از توزیع نرمال $\dot{\epsilon} \sim N(0, \hat{\sigma}^2)$ انتخاب شده است. مقادیری که به‌صورت تصادفی انتخاب می‌شوند با یک نقطه در بالای نماد نشان داده می‌شوند. روش SRI می‌تواند بایاس را با افزودن یک مرحله تکمیلی کاهش دهد که در آن، به نمره پیش‌بینی‌شده جداگانه یک جمله باقیمانده اضافه می‌گردد که از توزیع نرمال با میانگین صفر و واریانسی برابر با واریانس باقیمانده‌های رگرسیون اولیه، نمونه‌برداری می‌شود. میانگین صفر به‌عنوان یک شرط عدم بایاس اهمیت دارد و واریانس باید برابر با واریانس خطا تنظیم گردد. با درنظرگرفتن فرض‌های $E(\varepsilon) = 0, V(\varepsilon) = \hat{\sigma}^2 I$، توزیع ε_i به شرط x_i، برای تمامی مقادیر X صدق می‌کند که x_i نشان‌دهنده سطر i^{th} از ماتریس X است، همان‌طورکه در روابط (۱۱) و (۱۲) نشان داده شده است.

اگر $p(\varepsilon_i|x_i)$ نشان‌دهنده تابع چگالی احتمال شرطی ε_i به شرط x_i و $p(\varepsilon_i)$ نشان‌دهنده تابع چگالی احتمال غیرشرطی ε_i باشد، آن‌گاه:

$$E(\varepsilon_i|x_i) = \int \varepsilon_i p(\varepsilon_i|x_i) d\varepsilon_i$$
$$= \int \varepsilon_i p(\varepsilon_i) d\varepsilon_i \quad \text{رابطه ۱۱)}$$
$$= E(\varepsilon_i)$$
$$= 0$$

$$E({\varepsilon_i}^2|x_i) = \int {\varepsilon_i}^2 p(\varepsilon_i|x_i') d\varepsilon_i$$
$$= \int {\varepsilon_i}^2 p(\varepsilon_i) d\varepsilon_i \quad \text{رابطه ۱۲)}$$
$$= E({\varepsilon_i}^2)$$
$$= 0$$

در حالتی که ε_i و x_i مستقل باشند، آن‌گاه $p(\varepsilon_i|x_i') = p(\varepsilon_i)$. ما یک انحراف نرمال تصادفی را که با خطای استاندارد برآوردشده جریان رودخانه مقیاس‌دهی شده است، را به مدل اضافه کرده‌ایم.

۳. ۲. ۱. ۳. رگرسیون خطی بیزی

در BLR، رابطه رگرسیونی خطی نه با برآوردهای نقطه‌ای، بلکه به وسیله یک توزیع احتمال بیان می‌شود. فرض بر این است که پاسخ، y، از یک توزیع احتمال نمونه‌گیری می‌شود، نه این‌که به‌صورت یک مقدار منفرد محاسبه گردد (Kim & Lee, 2009). مدل BLR به‌شرح زیر است:

$$\dot{y} = \hat{\beta}_0 + X_{mis}\hat{\beta}_1 + \dot{\epsilon}, \quad \text{رابطه ۱۳)}$$

که در آن، $\dot{\epsilon} \sim N(0, \dot{\sigma}^2)$ و $\dot{\beta}_0, \dot{\beta}_1$ و $\dot{\sigma}^2$) همگی نمونه‌های تصادفی از توزیع پسین حاصل از داده‌ها استخراج می‌شوند. نمادگذاری ماتریسی در رابطه (۱۴) به‌صورت زیر بیان می‌شود:

$$y \sim N(X\beta, \sigma^2 I) \quad \text{رابطه ۱۴)}$$

در این‌جا، $y = (y_1, y_2, \dots, y_n)'$، $\beta = (\beta_1, \beta_2, \dots, \beta_k)'$ و X یک ماتریس $(N \times k)$ از N مشاهده و k متغیرهای مستقل $x_i = (x_{i1}, x_{i2}, \dots, x_{ik})$. علاوه بر این‌که خود متغیر پاسخ از یک توزیع احتمال مشتق می‌شود، پارامترهای مدل نیز انتظار می‌رود که از یک توزیع احتمال به‌دست آیند. احتمال پسین پارامترهای مدل وابسته به ورودی‌ها و خروجی‌های داده‌ی آموزشی است.

فرض که توزیع پیشین استاندارد غیراطلاع‌رسان برای پارامتر به‌صورت زیر باشد:

$$P(\beta, \sigma^2) \propto 1/\sigma^2 \quad \text{رابطه ۱۵)}$$

سپس توزیع احتمال پسین پارامترهای مدل به‌صورت زیر به‌دست می‌آید:

$$P(\beta, \sigma^2|y) = P(\beta, \sigma^2|y)P(\sigma^2|y)$$
$$\beta|\sigma^2, y \sim N\,(\widehat{\beta}, \sigma^2 V_\beta)$$
$$\sigma^2|y \sim Inv - Gamma\,(\frac{[n-k]}{2}, \frac{[n-k]s^2}{2}) \quad \text{رابطه ۱۶)}$$
$$\beta|y = t_{n-k}(\widehat{\beta}, s^2 V_\beta)$$

برآوردگر حداقل مربعات معمولی (OLSE) برای β با کمینه‌سازی عبارت $(y - X\beta)'(y - X\beta)$ نسبت به β مطابق رابطه (۱۷) به‌دست می‌آید:

$$\hat{\beta} = (X'X)^{-1}X'_y \quad \text{رابطه ۱۷)}$$

و یک برآوردگر برای σ^2 به‌صورت زیر به‌دست می‌آید:

$$\sigma^2 = \frac{1}{n-k}(y - X\hat{\beta})'(y - X\hat{\beta}) \quad \text{رابطه ۱۸)}$$

و واریانس OLS برای β برابر است با:

$$V_\beta = \sigma^2(X'X)^{-1} \quad \text{رابطه ۱۹)}$$

توزیع احتمال پسین هنگامی مناسب است که $n > k$ و $rank(X) = k$ باشد. الگوریتم از یک پارامتر ریج k استفاده می‌کند تا از مشکلات مربوط به ماتریس‌های تک‌ارزش جلوگیری کند. این عدد باید روی یک مقدار غیرمنفی و نزدیک به صفر ثابت شود، مثلاً k=۰/۰۰۰۱. برای برخی داده‌ها ممکن است مقدار k بزرگ‌تری لازم باشد. مقدار بزرگ‌تر k، مانند k=۰/۱، با احتمال بیش‌تری می‌تواند باعث بایاس به سمت صفر گردد و باید از آن پرهیز شود.

۳. ۲. ۱. ۴. درختان طبقه‌بندی و رگرسیون

روش CART که توسط Breiman *et al.* (1984) معرفی گردید، یکی از روش‌های برجسته در کلاس الگوریتم‌های یادگیری ماشین است. برای تقسیم نمونه، مدل‌های CART نیازمند متغیرهای پیش‌بین هستند و از نقاط برش در این متغیرهای پیش‌بین استفاده می‌شود. با استفاده از این نقاط برش، داده‌ها به زیرمجموعه‌های بزرگ‌تر و همگن‌تر تقسیم می‌شوند. با تکرار عملیات تقسیم روی هر دو زیرنمونه، یک سری انشعاب ایجاد می‌شود که در نهایت یک درخت دودویی شکل می‌گیرد (Erdal & Karakurt, 2013). هر گره در درخت دارای یک قاعده انشعاب‌دهی است که با کمینه‌کردن خطای نسبی (RE) تعیین می‌شود. این خطای نسبی در مسئله رگرسیون، نشان‌دهنده مجموع مربعات خطا برای آن انشعاب است:

رابطه ۲۰)
$$RE(d) = \sum_{l=0}^{L}(y_l - \bar{y}_L)^2 + \sum_{r=0}^{R}(y_r - \bar{y}_R)^2$$

به‌گونه‌ای‌که، y_l و y_r به‌ترتیب بخش‌های چپ و راست می‌باشند، که در هریک، L و R مشاهدات y، با میانگین‌های مربوطه $\bar{y}_L$ و $\bar{y}_R$ وجود دارد. قاعده تصمیم d، یک نقطه در متغیر برآوردگر x است که بخش‌های چپ و راست را مشخص می‌کند. سپس، قاعده بخش‌بندی که خطای برآورد را کمینه می‌کند، برای ساخت یک رأس درخت استفاده می‌گردد.

۳. ۲. ۱. ۵. رگرسیون خطی چندگانه با تخمین بوت‌استرپ

بوت‌استراپ یک تکنیک رایج برای سنجش تغییرپذیری با نمونه‌برداری مجدد از داده‌ها است (Chhabra *et al.*, 2017). این روش، هر آزمون یا متریک مبتنی بر نمونه‌گیری تصادفی با جایگزینی را اعمال می‌کند. مقدار تخمینی گم‌شده با استفاده از رابطه (۲۱) پیش‌بینی می‌شود:

رابطه ۲۱)
$$\dot{y} = \dot{\beta}_0 + X_{mis}\dot{\beta}_1 + \dot{\epsilon}$$

که در آن، $\dot{\epsilon} \sim N(0, \dot{\sigma}^2)$ و $\dot{\beta}_0, \dot{\beta}_1$ و $\dot{\sigma}^2$ برآوردهای حداقل مربعاتی هستند که پس از انتخاب یک نمونه بوت‌استرپ از داده‌های مشاهده‌شده محاسبه شده‌اند. این الگوریتم، با ترسیم یک نمونه بوت‌استرپ از بخش تکمیل‌شده داده‌ها، مقادیر جایگزین را محاسبه می‌کند و سپس با درنظرگرفتن این نمونه بوت‌استرپ به‌عنوان یک برداشت که تغییرپذیری نمونه‌گیری را در پارامترها برآورد می‌کند، حداقل مربعات را تخمین می‌زند (Heitjan & Rubin, 1990). در مقایسه با روش بیزی، رویکرد بوت‌استراپ از تجزیه چولسکی اجتناب می‌کند و نیازی به نمونه‌گیری از توزیع χ^2 ندارد.

۳. ۲. ۱. ۶. رگرسیون خطی چندگانه

پس از جایگزینی تمام مقادیر گم‌شده با روش‌های مختلف، مجموعه‌داده‌ها به‌طور کامل با استفاده از رگرسیون خطی چندگانه (MLR) تحلیل می‌شوند تا بهترین رویکردها برای مواجهه با داده‌های گم‌شده در مجموعه‌داده‌های جریان روزانه تعیین شوند. تحلیل رگرسیون یک تکنیک آماری است که رابطه بین حداقل دو متغیر کمّی و متغیرهای پیش‌بینی‌شده آن‌ها را بررسی می‌کند (van Loon & Laaha, 2015). مدل رگرسیون خطی چندگانه، یک روش آماری پرکاربرد در بسیاری از حوزه‌ها از جمله هیدرولوژی است (Campozano *et al.*, 2014; Carey & Paige, 2016). پارامترهای MLR به‌صورت زیر بیان می‌شوند:

رابطه ۲۲)
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i(\beta), i = 1, \ldots, N$$

که در آن، Y_i مقدار متغیر وابسته، β_0 تا β_k ضرایب ثابت ناشناخته، X_i مقدار متغیر مستقل و ε_i خطای تصادفی است.

۳. ۲. ۲. بررسی عملکرد روش‌ها

تأثیر روش‌های جایگزینی داده‌های گم‌شده بر مجموعه‌داده‌های جریان با استفاده از سه معیار عملکردی بررسی می‌گردد. برای ارزیابی روش‌هایی جایگزینی داده‌های گم‌شده، مقادیر ضریب تعیین تعدیل‌شده ($Adj\ R^2$)، خطای استاندارد (RSE) و میانگین درصد خطای مطلق (MAPE) محاسبه شده‌اند. سنجه خطا، میزان انحراف بین مقادیر برآوردشده و مقادیر مشاهده‌شده متناظر آن‌ها را اندازه‌گیری می‌کند. مقدار $Adj\ R^2$، مقدار R^2 است که با درنظرگرفتن تعداد متغیرهای مستقل در مدل تعدیل شده است. مقادیر $Adj\ R^2$، بین صفر تا ۱ قرار دارند و نشان‌دهنده قدرت رابطه بین مشاهدات و برآوردها هستند، به‌طوری‌که هرچه این مقدار بیش‌تر باشد، نشان‌دهنده عملکرد بهتر روش‌های برآورد است. اگر $Adj\ R^2$ به صفر نزدیک شود، عملکرد مدل نامطلوب یا ضعیف تلقی می‌شود. در مقابل، اگر مقادیر به یک نزدیک باشند، پیش‌بینی مدل عالی در نظر گرفته می‌شود (Mispan *et al.,* 2015; Rahman *et al.,* 2015). در همین‌حال، مقادیر کم‌تر برای RSE و MAPE نشان‌دهنده عملکرد بهتر روش‌های برآورد است. این آماره‌ها با استفاده از روابط (۲۳) تا (۲۵) محاسبه شده‌اند:

$$Adj\ R^2 = \bar{R}^2 = 1 - (1 - R^2)\left[\frac{n-1}{n-(k+1)}\right]$$ رابطه ۲۳)

$$RSE = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}{n-k-1}}$$ رابطه ۲۴)

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|x_i - \hat{x}_i|}{x_i}$$ رابطه ۲۵)

که در آن، x_i داده‌های مشاهده‌شده جریان، $\hat{x}_i$ مقدار برآوردشده، n حجم نمونه و k تعداد متغیرهای مستقل در معادله رگرسیون است.

۳. یافته‌های پژوهش و بحث

در این مطالعه، ارزیابی روش‌های MICE برای شناسایی بهترین تکنیک جایگزینی داده‌های گم‌شده در بازیابی داده‌های جریان روزانه رودخانه انجام شده است. ابتدا مدل‌ها بر روی مجموعه‌داده آموزشی دوره ۱۳۹۲-۱۳۹۱ تا ۱۴۰۱- ۱۴۰۰ که فاقد مقادیر گم‌شده بودند، آزمون شده‌اند. این دوره هیدرولوژیکی (۳۶۵۳ نقطه زمانی)، ۲۰ درصد از کل دوره (۱۸۲۶۲ نقطه زمانی) را شامل می‌گردد. فرایند شبیه‌سازی با روند زیر انجام شده است: یک مجموعه داده آموزشی با استفاده از نرخ‌های مختلف گمشدگی (پنج، ۱۰، ۱۵، ۲۰، ۲۵ و ۳۰ درصد) ایجاد و مقادیر گم‌شده مصنوعی با مقادیر جدیدی که با هریک از روش‌های MICE به‌دست آمدند، جایگزین گردید. خطا از تفاضل مقدار پیش‌بینی‌شده مدل آموزش دیده شده از مقدار پیش‌بینی‌شده مدل مرجع و داده‌های به‌دست‌آمده از روش جایگزینی مقادیر گم‌شده، محاسبه گردیده است. مدلی که با داده‌های آموزشی اصلی و داده‌های آزمونِ فاقد مقادیر گم‌شده آموزش دیده است، به‌عنوان مدل مرجع شناخته می‌شود. هرچه اختلاف بین مقادیر برآوردشده و مقادیر مشاهده‌شده کم‌تر باشد، مقادیر RSE و MAPE نیز کوچک‌تر خواهند بود. اگر مقدار برآوردشده با مقدار مشاهده‌شده مطابقت داشته باشد، مقدار $Adj\ R^2$ به عدد یک نزدیک خواهد بود. روش برازش بهینه براساس بیش‌ترین مقدار $Adj\ R^2$ و کم‌ترین مقادیر RSE و MAPE انتخاب خواهد شد. جدول (۲) خطاهای مدل پیش‌بینی را نشان می‌دهد، در حالی‌که جدول‌های (۳) تا (۵) نتایج انحراف را ارائه می‌کنند.

Table 2. Error of streamflow reference model in Kajo River basin

Station	Year	$Adj\ R^2$	RSE	MAPE
Pirsohrab	2011-2012 to 2021-2022	0.721	0.328	0.574
Chandokan	2011-2012 to 2021-2022	0.615	0.426	0.621

Table 3. The performance of six different percentages of missing data compared based on $Adj\ R^2$

Station	Methods	Missing data rate					
		5%	10%	15%	20%	25%	30%
Pirsohrab	PMM	0.701	0.692	0.683	0.672	0.661	0.701
	SRI	0.685	0.674	0.663	0.654	0.655	0.685
	BLR	0.711	0.703	0.695	0.687	0.679	0.711
	CART	0.723	0.719	0.715	0.711	0.707	0.723
	BOOT	0.625	0.621	0.621	0.622	0.623	0.625
Chandokan	PMM	0.679	0.668	0.657	0.646	0.635	0.679
	SRI	0.695	0.686	0.677	0.668	0.659	0.695
	BLR	0.705	0.697	0.689	0.681	0.673	0.705
	CART	**0.718**	**0.714**	**0.710**	**0.706**	**0.702**	**0.718**
	BOOT	0.623	0.622	0.622	0.623	0.624	0.623

Table 4. The performance of six different percentages of missing data compared based on RSE

Station	Methods	Missing data rate					
		5%	10%	15%	20%	25%	30%
Pirsohrab	PMM	0.378	0.382	0.386	0.390	0.394	0.378
	SRI	0.362	0.366	0.370	0.374	0.378	0.362
	BLR	0.348	0.352	0.356	0.360	0.364	0.348
	CART	0.311	0.315	0.319	0.323	0.327	0.311
	BOOT	0.392	0.396	0.400	0.404	0.408	0.392
Chandokan	PMM	0.374	0.378	0.382	0.386	0.390	0.374
	SRI	0.358	0.362	0.366	0.370	0.374	0.358
	BLR	0.344	0.348	0.352	0.356	0.360	0.344
	CART	0.307	0.311	0.315	0.319	0.323	0.307
	BOOT	0.388	0.392	0.396	0.400	0.404	0.388

Table 5. The performance of six different percentages of missing data compared based on MAPE

Station	Methods	Missing data rate					
		5%	10%	15%	20%	25%	30%
Pirsohrab	PMM	0.921	0.987	1.053	1.119	1.185	0.921
	SRI	0.815	0.874	0.933	0.992	1.051	0.815
	BLR	0.729	0.782	0.835	0.888	0.941	0.729
	CART	**0.437**	**0.488**	**0.539**	**0.590**	**0.641**	**0.437**
	BOOT	1.185	1.231	1.277	1.323	1.369	1.185
Chandokan	PMM	0.915	0.981	1.047	1.113	1.179	0.915
	SRI	0.809	0.868	0.927	0.986	1.045	0.809
	BLR	0.723	0.776	0.829	0.882	0.935	0.723
	CART	**0.431**	**0.482**	**0.533**	**0.584**	**0.635**	**0.431**
	BOOT	1.179	1.225	1.271	1.317	1.363	1.179

تحلیل شکاف نشان می‌دهد که کدام روش جایگزینی داده‌های گم‌شده از سازگاری بیش‌تری برخوردار است، به‌طوری‌که تفاوت کم‌تری بین نتایج مرحله آموزش و اعتبارسنجی وجود داشته باشد. از نتایج خلاصه‌شده در جدول‌های (۲) تا (۵) می‌توان مشاهده کرد که روش CART بیش‌ترین مقدار $Adj\ R^2$ و کم‌ترین مقادیر RSE و MAPE را تولید نموده است. در مقابل، روش BOOT بدترین روش جایگزینی برای داده‌های جریان روزانه در حوضه رودخانه کاجو بود که کم‌ترین مقدار $Adj\ R^2$ و بیش‌ترین مقادیر RSE و MAPE را داشت. از سوی دیگر، مقادیر $Adj\ R^2$ نشان دادند که تمامی روش‌های جایگزینی داده‌های گم‌شده نتایج قابل‌قبولی را ارائه می‌دهند با مقادیری نزدیک به یک و تفاوت کم‌تر از ۱۰ درصد نسبت به مجموعه آموزشی، در حالی‌که RSE با افزایش نرخ گمشدگی داده‌ها، مقادیر کمی پایین‌تر از مجموعه‌های آموزشی تولید می‌کند. هم‌چنین، MAPE بزرگی خطا را برحسب درصد اندازه‌گیری می‌کند و مقادیر آن بسته به میانگین اختلاف بین مقادیر واقعی مشاهده‌شده و مقادیر پیش‌بینی مدل، کمی نوسان دارد.

همان‌طورکه مشاهده می‌شود، دقت مدل با افزایش نرخ داده‌های گم‌شده کاهش نمی‌یابد. یک توضیح احتمالی برای افزایش کارایی با روش MICE این است که این روش می‌تواند با درنظرگرفتن روابط غیرخطی میان متغیرهای پیش‌بین،

از اطلاعات موجود استفاده بهینه‌تری ببرد (Islam Khan & Hoque, 2020). روش MICE به‌دلیل سادگی، استحکام، توانایی مدیریت چندخطی‌بودن و توزیع‌های اریب، و انعطاف‌پذیری برای تطابق با برهم‌کنش‌ها و روابط غیرخطی شناخته شده است (van Buuren & Groothuis-Oudshoorn, 2011). با افزایش نرخ داده‌های گم‌شده، خطا بین مدل مرجع و مدل‌های اعتبارسنجی که از جایگزینی داده‌های گم‌شده استفاده می‌کنند، بیش‌تر می‌شود. این نشان می‌دهد که وقتی مدل با داده‌های کامل (بدون مقادیر گم‌شده) آموزش می‌بیند، خطای آن بسیار ناچیز است. حتی اگر نرخ داده‌های گمشدگی تنها ۲۰ درصد بود، مدل آموزش‌دیده الگوی مربوط به ۸۰ درصد داده‌های آموزشی باقی‌مانده را دنبال کرد، نه الگوی ۲۰ درصد داده‌های گم‌شده را.

از نتایج به‌دست‌آمده، روش CART در مقایسه با چهار روش دیگر، یعنی PMM، SRI، BLR و BOOT عملکرد بهتری نشان داده است. همان‌طورکه در جدول (۳) مشاهده می‌شود، روش CART بالاترین مقدار $Adj\ R^2$ را به‌دست آورده است. از میان پنج روش، ضعیف‌ترین عملکرد مربوط به استفاده از روش BOOT بوده که در آن بدون ملاحظه نرخ‌های متفاوت گمشدگی، کم‌ترین مقدار $Adj\ R^2$ ثبت شده است. جدول‌های (۴) و (۵) به‌ترتیب شاخص عملکرد مربوط به RSE و MAPE را ارائه می‌دهند. بهترین عملکرد با کم‌ترین مقدار RSE مربوط به‌روش CART می‌باشد و پس از آن روش PMM قرار دارد. این یافته‌ها با مطالعات پیشین (Vezza *et al.,* 2010; Erdal & Karakurt, 2013; Karakurt *et al.,* 2013; Tyralis *et al.,* 2019; Hamzah *et al.,* 2022) نیز همسو است که در آن‌ها مدل CART از نظر واریانس تبیین‌شده بر سایر الگوریتم‌های طبقه‌بندی برتری نشان داده‌اند. نتایج پژوهش Erdal & Karakurt (2013) نیز نشان می‌دهد که مدل CART روش امیدبخش برای پیش‌بینی جریان ماهانه رودخانه بوده و در مقایسه با سایر مدل‌های ارزیابی‌شده، نتایج بهتری ارائه می‌دهد. در پژوهش Karakurt *et al.* (2013) نیز نتیجه گرفته شده است که یک مدل مبتنی بر طبقه‌بندی و رگرسیون، عملکرد بهتر (هرچند اندک) از شبکه عصبی مصنوعی (ANN) داشته است، به‌طوری‌که مقدار R^2 به‌ترتیب ۰/۸۹۹۸ و ۰/۸۹۴۲ بوده است. روش CART، هم‌چنین توسط Vezza *et al.* (2010) موردبررسی قرار گرفته است. در این پژوهش، از تحلیل واریانس یک‌طرفه برای برآورد واریانس تبیین‌شده توسط طبقه‌بندی CART استفاده شده و نتیجه ۶۹ درصد حاصل شده است. این یافته به این نتیجه‌گیری منجر می‌شود که روش CART یک روش طبقه‌بندی عالی است که توانایی شناسایی گروه‌های متمایز را هم از نظر پاسخ حوضه آبخیز در شرایط کم‌جریان و هم از نظر ویژگی‌های حوضه آبخیز دارد. مدل CART هم‌چنین می‌تواند معیارهای اهمیت متغیرها را ارائه دهد که این ویژگی، آن را از کلاس عمومی مدل‌های داده‌محور که تنها بر مدل‌سازی پیش‌بینانه متمرکز هستند، متمایز می‌کند (Tyralis *et al.,* 2019). به‌طور کلی، عملکرد برتر روش CART در این مطالعه به سازگاری ذاتی معماری درختی با ویژگی‌های آماری داده‌های جریان روزانه رودخانه کاجو شامل ماهیت به‌شدت غیرخطی، راست چوله و پُرنوسان بودن، مربوط است. زیرا CART، به گونه‌ای است که به هیچ‌یک از مفروضات پارامتریک (نرمال‌بودن، همگنی واریانس، خطی‌بودن رابطه) وابسته نمی‌باشد.

از سوی دیگر، روش BOOT ضعیف‌ترین عملکرد را نشان داد و هم‌چنین برای مجموعه‌داده‌های بزرگ، زمان‌برترین روش است. در همین حال، روش SRI که اغلب به‌عنوان یک رویکرد محافظه‌کارانه و ایمن برای برخورد با داده‌های گم‌شده در نظر گرفته می‌شود، با افزایش نرخ داده‌های گم‌شده، واریانس را کم‌تر از حد واقعی برآورد کرد. این اتفاق به این دلیل رخ می‌دهد که روش SRI می‌تواند نتایج غیرمنطقی یا غیرقابل‌قبول تولید کند. متغیرها در داده‌های جریان رودخانه معمولاً به بازه‌های مشخصی محدود می‌شوند (مانند مثبت‌بودن مقدار)، و روش SRI نمی‌تواند داده‌های گم‌شده را با درنظرگرفتن چنین محدودیت‌هایی بازسازی کند. به‌طور مشابه، روش جایگزینی BLR فرض می‌کند که خطای

تصادفی برای همه متغیرها در توزیع، میانگین یکسانی دارد که این امر منجر به ایجاد خطاهای بسیار کوچک یا بسیار بزرگ برای مقادیر جایگزین‌شده می‌شود. به‌طور کلی، روش BLR در مقایسه با رویکردهای PMM و CART عملکرد ضعیفی داشت. در مقابل، روش PMM تا سطح ۳۰ درصد داده‌های گم‌شده، تحت تأثیر قابل‌توجهی قرار نگرفت که احتمالاً به‌دلیل مکانیسم جایگزینی آن است که براساس مقادیر معتبر مشاهده‌شده در سایر بخش‌های داده صورت می‌گیرد. این روش مقادیر جایگزین‌شده‌ای تولید می‌کند که شباهت قابل‌توجهی به مقادیر واقعی دارند و از مفهوم استقراض از افرادی که داده‌های واقعی در اختیار دارند استفاده می‌کند (Schenker & Taylor, 1996). در نتیجه، گستره مقادیر جایگزین‌شده بین کمینه و بیشینه مقادیر مشاهده‌شده قرار می‌گیرد. جایگزینی داده‌ها خارج از محدوده داده‌های مشاهده‌شده صورت نمی‌گیرد و از این طریق، از بروز مشکلات مرتبط با جایگزینی‌های بی‌معنا اجتناب می‌شود (برای مثال، مقادیر غیرمثبت جریان رودخانه). با وجود این‌که Dong & Peng (2013) روش PMM را به‌عنوان یکی از بهترین و کاربردی‌ترین روش‌های جایگزینی برای متغیرهای پیوسته گم‌شده پیشنهاد کرده‌اند، این روش فاقد مبانی نظری بوده و به‌عنوان یک مسئله بهینه‌سازی فرمول‌بندی صریحی ندارد (Bertsimas *et al.*, 2018). عملکرد پنج روش جایگزینی از نظرMAPE با عملکرد آن‌ها در شاخص‌های $Adj\ R^2$ و RSE قابل مقایسه است. روش CART صرف‌نظر از هرگونه داده گم‌شده، با کم‌ترین مقادیر RSE و MAPE و بالاترین مقدار $Adj\ R^2$، عملکرد بهتری نسبت به سایر روش‌های موردمطالعه نشان داده است.

مدل‌ها در مرحله بعد با استفاده از داده‌های دوره هیدرولوژیکی ۱۳۵۲–۱۳۵۱ تا ۱۴۰۱–۱۴۰۰ برای هر دو ایستگاه آب‌سنجی حوضه رودخانه کاجو اعتبارسنجی شدند. جدول (۶) نتایج عملکرد کلی روش‌ها در بازسازی داده‌های دوره دوره هیدرولوژیکی ۱۳۵۲–۱۳۵۱ تا ۱۴۰۱–۱۴۰۰ را نمایش می‌دهد. براساس جدول (۶) روش CART بهترین عملکرد را داشت. در همین حال، روش BOOT بدترین روش جایگزینی برای داده‌های روزانه جریان رودخانه در حوضه رودخانه کاجو بود که کم‌ترین مقدار $Adj\ R^2$ و بیش‌ترین مقادیر RSE و MAPE را داشت. جدول (۶) هم‌چنین نشان می‌دهد که روش جایگزینیPMM در مقایسه با چهار روش دیگر، مقدار $Adj\ R^2$ بالاتر و مقادیر RSE و MAPE پایین‌تری دارد که آن را هم‌رده روشCART قرار می‌دهد.

پس از تکمیل مقادیر گم‌شده، در این مطالعه از مدل رگرسیون خطی چندگانه (MLR) برای تحلیل کل مجموعه‌داده استفاده شده است. از مدل MLR برای شناسایی بهترین رویکردهای مواجهه با داده‌های گم‌شده در زمانی که مقادیر جایگزین‌شده با مدل‌سازی ترکیب می‌شوند، استفاده شده است. جدول (۷) عملکرد هر پنج روش جایگزینی را در ترکیب با مدل MLR برای پیش‌بینی نرخ جریان رودخانه در حوضه رودخانه کاجو از سال هیدرولوژیکی ۱۳۵۲–۱۳۵۱ تا ۱۴۰۱–۱۴۰۰ را نشان می‌دهد. براساس جدول (۷)، روش CART-MLR بهترین عملکرد را ارائه داده است، درحالی‌که روش BOOT-MLR در میان پنج روش ارزیابی‌شده ضعیف‌ترین عملکرد را داشته است. اگرچه روش PMM-MLR در مقایسه با روش‌های SRI-MLR، BLR-MLR و BOOT-MLR عملکرد کمی بهتر داشت، اما روش CART-MLR بر سایر رویکردها برتری نشان داد.

یافته‌های این مطالعه نشان می‌دهد که روش CART در ترکیب با MLR به‌طور معناداری از سایر روش‌های آزمون‌شده بهتر عمل کرده است. این نشان می‌دهد که خطای ناشی از روش CART در مقایسه با روش‌های PMM، SRI، BLR و BOOT به نسبت کم‌تر بود، زیرا نرخ خطا بازتابی از نرخ داده‌های گم‌شده می‌باشد. CART به‌عنوان یک مدل ترکیبی مبتنی بر درخت، می‌تواند دقت خود را با تولید مجموعه‌داده‌های متعدد مشابه و توسعه مدل‌های گوناگون با سوگیری کم‌تر، به‌طور معقولانه‌ای افزایش دهد و سپس آن‌ها را در ساخت یک مدل ترکیبی با عملکرد بالاتر ادغام کند

(Erdal & Karakurt, 2013: Tyralis *et al.,* 2019). این یافته‌ها با نتایج گزارش‌شده در پیشینه پژوهش همسو است و توصیه‌های مربوط به‌روش جایگزینی CART را تأیید می‌کند (De'ath & Fabricius, 2000; Erdal & Karakurt, 2013; Karakurt *et al.,* 2013). در نهایت، این شبیه‌سازی‌ها نشان می‌دهند که روش CART در ترکیب با MLR، بهترین روش جایگزینی داده‌های گم‌شده برای بازسازی داده‌های جریان رودخانه است. شکل (۶) هیدروگراف‌های مشاهده‌شده و بازسازی‌شده با CART-MLR را در طول دوره ناقص هیدرولوژیکی (۱۳۵۲-۱۳۵۱ تا ۱۳۹۰-۱۳۹۰) در دو ایستگاه آب‌سنجی پیرسهراب و چندوکان نشان می‌دهد. این سری‌های زمانی هیدرولوژیکی، برای هرگونه تحلیل‌های هیدرولوژیکی حوضه آبریز رودخانه کاجو توصیه می‌گردد.

Table 6. Performance criteria values for five imputation methods in Kajo River basin

Station	Methods	Performance criteria		
		$Adj\ R^2$	RSE	MAPE
Pirsohrab	PMM	0.702	0.435	0.521
	SRI	0.623	0.487	1.301
	BLR	0.655	0.472	1.248
	CART	0.741	0.379	0.462
	BOOT	0.569	0.524	1.335
Chandokan	PMM	0.688	0.451	0.538
	SRI	0.612	0.493	1.275
	BLR	0.645	0.476	1.210
	CART	0.728	0.392	0.495
	BOOT	0.558	0.518	1.312

Table 7. The results for MLR when combined with imputation methods

Station	Methods	Performance criteria		
		$Adj\ R^2$	RSE	MAPE
Pirsohrab	PMM-MLR	0.635	0.493	0.742
	SRI-MLR	0.590	0.529	1.785
	BLR-MLR	0.568	0.547	1.155
	CART-MLR	**0.785**	**0.472**	**0.583**
	BOOT-MLR	0.305	0.558	1.892
Chandokan	PMM-MLR	0.632	0.496	0.751
	SRI-MLR	0.588	0.531	1.793
	BLR-MLR	0.565	0.549	1.162
	CART-MLR	**0.780**	**0.475**	**0.588**
	BOOT-MLR	0.310	0.560	1.880

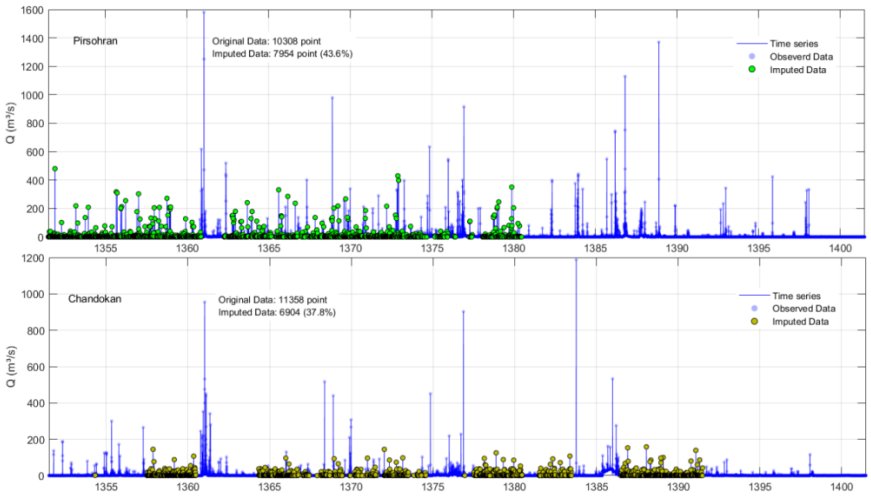

Figure 6. Observed (OBS) and simulated (SIM) hydrographs in Kajo River basin

۴. نتیجه‌گیری

داده‌های گم‌شده همواره به تفسیر نادرست خروجی‌های آماری منجر می‌شوند، بنابراین روش مورداستفاده برای پُرکردن شکاف‌ها در یک مجموعه‌داده باید با دقت موردتوجه قرار گیرد. چندین روش برای مدیریت داده‌های گم‌شده در پیشینه پژوهش ارائه شده است و انتخاب رویکرد مناسب، از جمله الگوی داده‌های گم‌شده و سازوکار گمشدگی، هنوز مبهم باقی مانده است. روش‌های جایگزینی داده‌ها، ازدست‌رفتن اطلاعات را کاهش داده‌اند؛ اطلاعاتی که در غیر این صورت می‌توانست به بروندادهای بهینه‌نشده و نتیجه‌گیری‌های گمراه‌کننده، از جمله برآورد خطر یک رویداد شدید، منجر شود. هدف این پژوهش، ارزیابی عملکرد روش انتساب چندگانهٔ جایگزینی مبتنی بر معادلات زنجیره‌ای (MICE) را در پیش‌بینی و بازسازی مقادیر جریان روزانه رودخانه در حوضه آبریز رودخانه کاجو بود. رودخانه کاجو یکی از بسترهای سیلابی مهم در جنوب‌شرق ایران شناخته می‌شود. داده‌های این پژوهش شامل جریان روزانه رودخانه در دو ایستگاه هیدرومتری واقع بر روی رودخانه کاجو شامل ایستگاه پیرسهراب و ایستگاه چندوکان بود. دوره آماری مورداستفاده هر دو ایستگاه از سال آبی ۱۳۵۱/۱۳۵۲ تا ۱۴۰۱/ ۱۴۰۰ بود. برای بررسی و اعتبارسنجی کارایی رویکرد MICE در مدیریت داده‌های گم‌شده جریان، از داده‌های تاریخی روزانه کامل جریان در بازه سال‌های هیدرولوژیکی ۱۳۹۲–۱۳۹۱ تا ۱۴۰۱– ۱۴۰۰ استفاده شده است. سپس از ارزیابی و اعتبارسنجی کارایی رویکرد MICE در مدیریت داده‌های گم‌شده با استفاده از داده‌های کامل جریان رودخانه، این روش‌ها همراه با رگرسیون خطی چندگانه (MLR) برای بازسازی کُل مقادیر گم‌شده جریان روزانه به‌کار گرفته شد.

نتایج این مطالعه نشان داد که روش CART، صرف‌نظر از درصد مقادیر گم‌شده، به‌طور پیوسته برتر بوده است. هر سه شاخص عملکرد توافق داشتند که روش CART در زمره بهترین‌ها با مقدار $Adj\ R^2$ بالاتر و مقادیر RSE و MAPE پایین‌تر در مقایسه با سایر روش‌های معرفی‌شده توسط الگوریتم MICE قرار دارد. نتایج هم‌چنین نشان داد که روش CART کم‌ترین تفاوت را بین مدل مرجع و مدل پیش‌بینی با جایگزینی داده‌های گم‌شده ایجاد می‌نماید. در نتیجه، بهترین نتایج با پردازش داده‌های جریان گم‌شده با استفاده از CART در ترکیب با MLR به‌دست آمده است. در نهایت، این پژوهش به پُرکردن دقیق مجموعه‌داده‌های گم‌شده جریان روزانه رودخانه کمک می‌کند. این مطالعه بر پایه عملکرد MICE به‌عنوان مدل شرطی برای جایگزینی داده‌ها در برآورد سوابق جریان گم‌شده قرار دارد. با توجه به موفقیت روش CART در مدیریت پیچیدگی‌های آماری داده‌های جریان (چولگی شدید، نوسانات زیاد و مقادیر پرت)، به‌عنوان گزینه‌ای قابل‌قبول در مطالعات مشابه در مناطق خشک و سیل‌خیر پیشنهاد می‌گردد. باوجود نتایج رضایت‌بخش، MICE با محدودیت روش‌شناسی مهمی روبه‌رو بوده که فقدان توجیه نظری از جمله است. به‌گونه‌ای‌که این رویه بیش‌تر مبتنی بر کاربرد عملی و موفقیت تجربی می‌باشد تا مبانی نظری محکم. بنابراین پیشنهاد می‌گردد در مطالعات آتی، قابلیت روش‌های جدیدتر مانند یادگیری عمیق نیز آزموده شود.

یافته‌های این پژوهش پیامدهای مستقیمی برای مدیریت یکپارچه منابع آب و کنترل سیلاب در حوضه کاجو دارد. بازسازی دقیق و قابل اعتماد داده‌های جریان روزانه، پایه‌ای اساسی برای کالیبراسیون و اعتبارسنجی مدل‌های هیدرولوژیکی و هیدرولیکی، تخمین دقیق پارامترهای طراحی و مدیریت تنظیم سد زیردان، برآورد احتمالاتی سیلاب و تدوین طرح‌های بهره‌برداری بهینه از مخزن فراهم می‌کند. دست‌یابی به یک سری زمانی کامل و باکیفیت، امکان تحلیل‌های فرکانسی، شناسایی روندهای بلندمدت و ارزیابی اثرات تغییر اقلیم بر رژیم هیدرولوژیکی حوضه را نیز میسر می‌سازد.

۵. مشارکت نویسندگان

فاطمه گنجی‌گوهری: گردآوری داده‌ها و مدیریت داده‌ها، تحلیل و نوشتن پیش‌نویس؛

حمید نظری‌پور (نویسنده مسئول): طراحی روش‌شناسی، تحلیل، ویرایش نهایی و نظارت بر پژوهش؛

محمدرضا پودینه و محسن حمیدیان‌پور: تحلیل داده‌ها و بازبینی محتوا؛

علیرضا قائمی: پیش‌پردازش داده‌ها؛

رضا تیموری: تحلیل‌های تکمیلی و ویرایش نهایی مقاله.

۶. تشکر و قدردانی

نویسندگان صمیمانه از Stef van Buuren، استاد تحلیل آماری داده‌های ناقص در دانشگاه اوترخت، به پاس ابداع الگوریتم MICE و به اشتراک‌گذاری سخاوتمندانه و عمومی نسخه نرم‌افزاری آن، تشکر و قدردانی می‌کنند. این اثر، تحت حمایت مادی بنیاد ملی علم ایران (INSF) برگرفته شده از طرح شماره ۴۰۴۴۳۳۴ انجام شده است.

۷. تعارض منافع

هیچ‌گونه تعارض منافع توسط نویسندگان وجود ندارد.

۸. منابع

Adeloye, A.J. (1996). An opportunity loss model for estimating the value of streamflow data for reservoir planning. *Water Resources Management*, 10 (1), 45-79.

Ahmat Zainuri, N., Aziz Jemain, A., & Muda, N. (2015). A comparison of various imputation methods for missing values in air quality data. *Sains Malaysiana*, 44 (3), 449-456.

Ahn, K.H. (2021). Streamflow estimation at partially gaged sites using multiple-dependence conditions via vine copulas. *Hydrology and Earth System Sciences*, 25 (8), 4319-4333.

Aryanmanesh, J., Nazaripour, H., Mahmoodi, P., & Khosravi, P. (2024). Reconstruction of Missing Daily Streamflow Data using the MissForest Algorithm in Southern Baluchestan Basin, Iran. *J Watershed Manage Res*., *15*(2), 49-64.

Bertsimas, D., Pawlowski, C., & Zhuo, Y.D. (2018). From predictive methods to missing data imputation: an optimization approach. *Journal of Machine Learning Research*, 18 (2018), 1-39.

Breiman, L., *et al*. (1984). *Classification and regression trees*. New York: Wadsworth Publishing.

Campozano, L., *et al*. (2014). Evaluation of infilling methods for time series of daily precipitation and temperature: the case of the Ecuadorian Andes. *Maskana*, 5 (1), 99-115.

Carey, A.M., & Paige, G.B. (2016). Ecological site-scale hydrologic response in a semiarid rangeland watershed. *Rangeland Ecology and Management*, 69 (6), 481-490.

Chhabra, G., Vashisht, V., & Ranjan, J. (2017). A comparison of multiple imputation methods for data with missing values. *Indian Journal of Science and Technology*, 10 (19), 1-7.

De'ath, G., & Fabricius, K.E. (2000). Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, 81 (11), 3178-3192.

Devineni, N., *et al*. (2013). A tree-ring-based reconstruction of Delaware River basin streamflow using hierarchical Bayesian regression. *Journal of Climate*, 26 (12), 4357-4374.

Donders, A.R.T., *et al*. (2006). Review: a gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, 59 (10), 1087-1091.

Dong, Y., & Peng, C.-Y.J. (2013). Principled missing data methods for researchers. *SpringerPlus*, 2 (1), 1-17.

Erdal, H.I., & Karakurt, O. (2013). Advancing monthly streamflow prediction accuracy of CART models using ensemble learning paradigms. *Journal of Hydrology*, 477 (2013), 119-128.

Feizi, H., & Sattari, M. T. (2026). Streamflow Forecasting Based on PatchTST, LSTM, and Ensemble Learning Approaches. *Water Resources Management*, *40*(1), 44.

Gao, Y. (2017). *Dealing with missing data in hydrology - data analysis of discharge and groundwater time-series in Northeast Germany*. Germany: Freie Universität Berlin.

Gill, M.K., *et al*. (2007). Effect of missing data on performance of learning algorithms for hydrologic predictions: implications to an imputation technique. *Water Resources Research*, 43 (7), 1-12.

Gires, A., Tchiguirinskaia, I., & Schertzer, D. (2021). Infilling missing data of binary geophysical fields using scale invariant properties through an application to imperviousness in urban areas. *Hydrological Sciences Journal*, 66 (7), 1197-1210.

Hamzah, F.B., *et al*. (2020). Imputation methods for recovering streamflow observation : a methodological review. *Cogent Environmental Science*, 6 (1), 21.

Hamzah, F.B., *et al*. (2021). A comparison of multiple imputation methods for recovering missing data in hydrological studies. *Civil Engineering Journal*, 7 (9), 1608-1619.

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.

Harvey, C.L., Dixon, H., & Hannaford, J. (2012). An appraisal of the performance of data-infilling methods for application to daily mean river flow records in the UK. *Hydrology Research*, 43 (5), 618-637.

Heitjan, D.F., & Rubin, D.B. (1990). Inference from coarse data via multiple imputation with application to age heaping. *Journal of the American Statistical Association*, 85 (410), 304-314.

Islam Khan, S., & Hoque, A.S.M.L. (2020). SICE: an improved missing data imputation technique background and related works. *Journal of Big Data*, 7 (1), 37.

Jamil, J.M. (2012). *Partial least squares structural equation modelling with incomplete data: an investigation of the impact of imputation methods*. The University of Bradford.

Johnston, C.A. (1999). *Development and evaluation of infilling methods for missing hydrologic and chemical watershed monitoring data*. Virginia Polytechnic Institute and State University.

Kamaruzaman, I.F., Wan Zin, W.Z., & Mohd Ariff, N. (2017). A comparison of method for treating missing daily rainfall data in Peninsular Malaysia. *Malaysian Journal of Fundamental and Applied Sciences*, 13 (4), 375-380. (Special Issue on Some Advances in Industrial and Applied Mathematics).

Karakurt, O., *et al*. (2013). Comparing ensembles of decision trees and neural networks for one-day-ahead stream flow predict. *Scientific Research Journal*, 1 (15), 43-54.

Kim, S.U., & Lee, K.S. (2009). Regional low flow frequency analysis using Bayesian regression and prediction at ungauged catchment in Korea. *KSCE Journal of Civil Engineering*, 14 (1), 87-98.

Little, R.J.A., & Rubin, D.B. (2002). Statistical analysis with missing data, hlm. 2nd Edisi. Hoboken, New Jersey: John Wiley & Sons, Inc.

Mispan, M.R., *et al*. (2015). Missing river discharge data imputation approach using artificial neural network. *ARPN Journal of Engineering and Applied Sciences*, 10 (22), 10480-10485.

Mohamad Hamzah, F., Mohd Yusoff, S.H., & Jaafar, O. (2019). L-moment-based frequency analysis of high-flow at the Sungai Langat, Kajang, Selangor, Malaysia. *Sains Malaysiana*, 48 (7), 1357-1366. L-Moment-Based.

Müller, K.-R., *et al.* (1997). Predicting time series with support vector machines. In: W. Gerstner, *et al.,* (eds) *Artificial Neural Networks — ICANN'97. ICANN 1997. Lecture Notes in Computer Science*. Berlin Heidelberg: Springer.

Mwale, F.D., Adeloye, A.J., & Rustum, R. (2012). Infilling of missing rainfall and streamflow data in the Shire River basin, Malawi - A self organizing map approach. *Physics and Chemistry of the Earth*, 50-52 (2012), 34-43.

Nekoeeyan, M. A., Radmanesh, F., & Ahmadi, F. (2022). Prediction of Monthly Streamflow Using Shannon Entropy and Wavelet Theory Approaches (Case study: Maroon River). *Water and Irrigation Management*, *12*(1), 15-31.

Nor, S.M.C.M., *et al*. (2020). A comparative study of different imputation methods for daily rainfall data in east-coast Peninsular Malaysia. *Bulletin of Electrical Engineering and Informatics*, 9 (2), 635-643.

Norazizi, N.A.A., & Deni, S.M. (2019). Comparison of Artificial Neural Network (ANN) and other imputation methods in estimating missing rainfall data at Kuantan Station. *Soft Computing in Data* Science, *5th International Conference, SCDS 2019*, hlm, Singapore. Iizuka, Japan: Springer, 298-308.

Plaia, A., & Bondì, A.L. (2006). Single imputation method of missing values in environmental pollution data sets. *Atmospheric Environment*, 40 (38), 7316-7330.

Rahman, N.F.A., *et al*. (2015). Semi distributed hydro climate model; The Xls2NCascii program approach for weather generator. *ARPN Journal of Engineering and Applied Sciences*, 10 (15), 6619-6622.

Regonda, S.K., *et al*. (2013). Short-term ensemble streamflow forecasting using operationally-produced single-valued streamflow forecasts-A Hydrologic Model Output Statistics (HMOS) approach. *Journal of Hydrology*, 497 (2013), 80-96.

Schenker, N., & Taylor, J.M.G. (1996). Partially parametric techniques for multiple imputation. *Computational Statistics and Data Analysis*, 22 (4), 425-446.

Schmitt, P., Mandel, J., & Guedj, M. (2015). A comparison of six methods for missing data imputation. *Journal of Biometrics & Biostatistics*, 6 (1), 1-6.

Semiromi, M.T., & Koch, M. (2019). Reconstruction of groundwater levels to impute missing values using singular and multichannel spectrum analysis: application to the Ardabil Plain, Iran. *Hydrological Sciences Journal*, 64 (14), 1711-1726.

Su, Y.-S., *et al*. (2011). Multiple imputation with diagnostics (mi) in R: opening windows into the black box. *Journal of Statistical SoftwareSoftware*, 45 (2), 31.

Tencaliec, P., *et al*. (2015). Reconstruction of missing daily streamflow data using dynamic regression models. *Water Resources Research, American Geophysical Union*, 51 (12), 9447-9463.

Tencaliec, P. (2017). *Developments in statistics applied to hydrometeorology: imputation of streamflow data and semiparametric precipitation modeling*. Universite Grenoble Alpes.

Tyralis, H., Papacharalampous, G., & Langousis, A. (2019). A brief review of random forests for water scientists and practitioners and their recent history in water resources. *Water*, 11 (5), 1-37.

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16 (3), 219-242.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice : Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45 (3), 1-67.

Van Loon, A.F., & Laaha, G. (2015). Hydrological drought severity explained by climate and catchment characteristics. *Journal of Hydrology*, 526, 3-14.

Vezza, P., *et al*. (2010). Low flows regionalization in north-western Italy. *Water Resources Management*, 24 (14), 4049-4074.

White, I.R., & Wood, A.M. (2011). Multiple imputation using chained equations : issues and guidance for practice. *Statistics in Medicine*, 30 (4), 377-399.

Widaman, K.F. (2006). Missing Data: what to do with or without them. *Monographs of the Society for Research in Child Development*, 71 (1), 210-211.

Zhao, Y., & Long, Q. (2016). Multiple imputation in the presence of high-dimensional data. *Statistical Methods in Medical Research*, 25 (5), 2021-2035.

Zvarevashe, W., Krishnannair, S., & Sivakumar, V. (2019). Analysis of rainfall and temperature data using ensemble empirical mode decomposition. *Data Science Journal*, 18 (1), 1-9.