Sari Agriculture Sciences and Natural Resources University

Journal of Watershed Management Research, Vol. 15, Issue 2, 2024 p: 49-64

OPEN ACCESS

Research Paper

Reconstruction of Missing Daily Streamflow Data using the MissForest Algorithm in Southern Baluchestan Basin, Iran

Javad Aryanmanesh[1], Hamid Nazaripour[2], Peyman Mahmoodi[3], Parviz Khosravi[4]

1- M.Sc Student, Department of physical Geography, University of Sistan and Baluchestan, Zahedan, Iran
2- Assistant Professor, Department of physical Geography, University of Sistan and Baluchestan, Zahedan, Iran (Corresponding author: h.nazaripour@gep.usb.ac.ir)
3- Associate Professor, Department of physical Geography, University of Sistan and Baluchestan, Zahedan, Iran
4- Master of information technology, Iran Meteorological Organization

Received: 20 November, 2023 Accepted: 9 March, 2024

Extended Abstract

Background: Long-term hydrometeorological variables can be used for planning and managing water resources at the basin level using different physical models, such as hydrological and hydraulic models. However, such variables are often accompanied by missing data, which makes analysis difficult or sometimes impossible. Data gaps cause problems in interpretation, model calibration, and biased statistics. In this study, the validity of a non-parametric random learning machine algorithm, called MissForest, has been evaluated to fill the gap of daily streamflow series in a region with scarce data and strong climate variability.

Methods: The daily streamflow data in the gauge stations of the Southern Baluchestan catchment were analyzed in a long-term hydrological period (09/23/1972 to 09/22/2018). First, the missingness percentage was selected based on a conventional criterion (less than 50%) as an acceptable ratio of the missing rate in the streamflow data, followed by investigating the mechanisms and patterns of the missing data. Accordingly, the number of gauge stations was reduced to seven samples. Then, the temporal distribution of the missing daily streamflows during the months of the year and the relative frequency of gap length were investigated during the period. Next, the performance of the missing data reconstruction algorithm was challenged with two different artificial missing data scenarios. Two types of artificial gaps were generated, namely a) Removed contiguous segments: at each gauge only a segment (having lengths of 7, 14, 21, 30, 60, 180, and 365 days) was randomly removed from the entire record (1972–2018); b) Removed single data points: observed values (30, 60, 90, 120, 180, and 365 days) were randomly removed from the entire record (1972–2018) at each of the gauges. *MissForest* was applied to fill the gaps contained in the records together with the artificial gaps. Our analysis includes reconstructions of the 1972–2018 period at each of the streamflow gauges. Finally, the performance of MissForest in infilling daily streamflow data was tested by comparing the filled series with the observed data using goodness-of-fit (GoF) indicators, coefficient of determination (R^2), the percent bias (PBIAS), and the Kling-Gupta efficiency (KGE).

Results: The MissForest algorithm generally performed satisfactorily, allowing for accurately and reliably simulating lost data quickly and automatically. The performance of the MissForest algorithm is highly dependent on the number of predictor records, record length, and streamflow type. Finally, the reconstruction of real gaps in streamflow data was possible by applying this intelligent algorithm. The river flow time series were simulated with the natural flow regime with good performance; however, this performance dropped slightly for flow rate changes as a result of water storage and diversion for irrigation, especially downstream of dams. The performance of this algorithm in filling the daily time series of flow with severe changes in the flow regime, such as peak discharge, was not evaluated optimally. This drop in performance is more related to the hydroclimatic conditions of the studied watershed than the structure of the algorithm. The reconstructed hydrographs allow for analyzing flow variability and their interaction with key climate variables.

Conclusion: The MissForest algorithm is introduced as one of the imputation methods based on machine learning with high credibility and performance in reconstructing the missing data of the daily streamflow. It can also be used automatically and intelligently in the reconstruction of the statistical defects of the river flow in the scale used daily. Future studies are suggested to analyze the effects of different watersheds with specific hydro-physical-climatic characteristics on the

[DOI: 10.61186/jwmr.15.2.49]

performance of the MissForest algorithm. The other issues that need to be addressed in future studies include the investigation of the proposed method of this study in other climatic and geographical regions, the sensitivity measurement to the rainfall and flow regime, and finally, the investigation of its performance compared to other common methods.

Keywords: Goodness of fit, Machine learning, MissForest algorithm, Missing data, Streamflow

How to Cite This Article: 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. DOI: 10.61186/jwmr.15.2.49

[DOI: 10.61186/jwmr.15.2.49]

دانشگاه علوم کشاورزی و منابع طبیعی ساری

مقاله پژوهشی

بازسازی داده‌های گمشده جریان روزانه رودخانه با استفاده از الگوریتم جنگل گمشده در حوزه بلوچستان جنوبی، ایران

جواد آریان‌منش[۱]، حمید نظری‌پور[۲]، پیمان محمودی[۳] و پرویز خسروی[۴]

۱- دانشجوی کارشناسی ارشد، گروه جغرافیای طبیعی، دانشکده جغرافیا و برنامه‌ریزی محیطی، دانشگاه سیستان و بلوچستان، زاهدان، ایران
۲- استادیار، گروه جغرافیای طبیعی، دانشکده جغرافیا و برنامه‌ریزی محیطی، دانشگاه سیستان و بلوچستان، زاهدان، ایران،
(نویسنده مسوول: h.nazaripour@gep.usb.ac.ir)
۳- دانشیار، گروه جغرافیای طبیعی، دانشکده جغرافیا و برنامه‌ریزی محیطی، دانشگاه سیستان و بلوچستان، زاهدان، ایران
۴- کارشناس ارشد فناوری اطلاعات، سازمان هواشناسی کشور

تاریخ دریافت: ۱۴۰۲/۸/۲۹ تاریخ پذیرش: ۱۴۰۲/۱۲/۱۹
صفحه: ۴۹ تا ۶۴

چکیده مبسوط

مقدمه و هدف: سری‌های زمانی کامل هیدرولوژیکی برای مدیریت و مدل‌سازی منابع آب و انرژی در یک اقلیم در حال تغییر حیاتی هستند. با این‌حال، چنین متغیرهایی اغلب با داده‌های گمشده همراه هستند، که فرایند تجزیه و تحلیل را دشوار و یا گاهی غیرممکن می‌کند. شکاف‌های داده باعث مشکلاتی در تفسیر، واسنجی ناکارآمد مدل و آماره‌های اُریب‌دار می‌شوند. در این بررسی، اعتبار یک الگوریتم ماشین یادگیری تصادفی غیرپارامتری که جنگل گمشده (*MissForest*) نام دارد برای پرکردن شکاف سری‌های زمانی جریان روزانه در منطقه‌ای با داده کمیاب و تغییرپذیری اقلیمی قوی، ارزیابی گردیده است.

مواد و روش‌ها: داده‌های جریان روزانه در ایستگاه‌های جریان‌سنجی حوزه آبریز بلوچستان جنوبی در یک دوره طولانی‌مدت هیدرولوژیکی (۱۹۷۲/۰۹/۲۳ تا ۲۰۱۸/۰۹/۲۲) مورد بررسی قرار گرفته است. منطقه مورد مطالعه این پژوهش (حوزه آبریز بلوچستان جنوبی) از مجموعه حوزه آبریز خلیج فارس و دریای عمان بوده و با حدود بین سدیج و مرکز پاکستان شناخته می‌شود. درصد گمشدگی بر اساس یک معیار قراردادی (کمتر از ۵۰ درصد) به‌عنوان نسبت قابل‌قبول از نرخ گمشدگی در داده‌های جریان انتخاب و سپس مکانیسم‌ها و الگوهای گمشدگی داده‌ها تعیین گردیده است. بر این اساس، تعداد ایستگاه‌های جریان‌سنجی از ۱۱ به ۷ نمونه کاهش یافته است. سپس توزیع زمانی جریان‌های روزانه گمشده در طول ماه‌های سال و فراوانی نسبی طول گمشدگی در کل دوره مورد بررسی قرار گرفته است. در ادامه، عملکرد الگوریتم بازسازی داده‌های گمشده با دو سناریوی متفاوت داده گمشده مصنوعی به چالش کشیده شده است. برای این‌منظور، دو نوع شکاف مصنوعی در قسمت داده‌های کامل ایجاد شده است. الف) در هر ایستگاه جریان‌سنجی یک بخش از داده‌ها (با طول ۷، ۱۴، ۲۱، ۳۰، ۶۰، ۱۸۰ و ۳۶۵ روز) به‌طور تصادفی از کل دوره حذف شده است. ب) نقاط داده منفرد شامل مقادیر مشاهده شده روزهای (۳۰، ۶۰، ۹۰، ۱۲۰، ۱۸۰ و ۳۶۵) به‌طور تصادفی از کل دوره (۲۰۱۸-۱۹۷۲) حذف شده‌اند. الگوریتم جنگل گمشده برای پُرکردن شکاف‌های مصنوعی اجرا و سپس اعتبارسنجی الگوریتم در پُرکردن داده‌های گمشده جریان روزانه با مقایسه سری‌های پُرشده با داده‌های مشاهده شده، از طریق آزمون‌های سه‌گانه نیکویی برازش (*GoF*) شامل ضریب تعیین (R2)، درصد بایاس یا اریب (PBIAS) و معیار کلینگ- کوپتا (KGE) تست شده است. علاوه بر آن، برخی کنترل‌ها در عملکرد الگوریتم جنگل گمشده جهت حساسیت‌سنجی انجام شده است. به این مفهوم که الگوریتم جنگل گمشده با درصدهای مختلف از گمشدگی داده در ایستگاه هدف (۵٪، ۱۰٪، ۱۵٪، ۲۰٪، ۲۵٪ و ۳۰٪) و همچنین تعداد رکوردهای پیش‌بینی کننده جریان ایستگاه هدف، آزمایش شده است.

یافته‌ها: نتایج نشان داد که به‌طور کلی الگوریتم جنگل گمشده عملکرد رضایت‌بخش و خوبی داشته و امکان شبیه‌سازی دقیق و مطمئن داده‌های از دست رفته را به‌سرعت و به‌صورت خودکار فراهم می‌آورد. عملکرد الگوریتم جنگل گمشده به‌شدت تابعی از تعداد رکوردهای پیش‌بینی کننده، طول رکورد و نوع جریان رودخانه می‌باشد. عملکرد الگوریتم جنگل گمشده به درصد گمشدگی داده‌های ایستگاه هدف حساس و به تعداد رکوردهای پیش‌بینی کننده بی‌تفاوت بوده است. با افزایش درصد گمشدگی داده‌ها، عملکرد الگوریتم جنگل گمشده به‌طور قابل ملاحظه کاهش یافته است. علاوه بر آن، این الگوریتم گمشدگی‌های کوتاه‌مدت را نسبت به گمشدگی‌های طولانی‌مدت، دقیق‌تر برآورد می‌کند. عملکرد الگوریتم جنگل گمشده به تعداد رکوردهای پیش‌بینی کننده حساس نمی‌باشد. این وضعیت، به ماهیت هیدروفیزیوگرافی زیرحوضه‌های آبریز و موقعیت ایستگاه‌های آب‌سنجی مربوط می‌شود. تنها در صورتی عملکرد الگوریتم جنگل گمشده برای یک ایستگاه خاص با افزایش رکوردهای پیش‌بینی کننده بهبود می‌یابد که ایستگاه‌های اهداءگر در حوضه آبریز مشترک با ایستگاه هدف قرار داشته باشند در نهایت، بازسازی شکاف‌های واقعی در داده‌های جریان از طریق اعمال این الگوریتم هوشمند ممکن گردید. سری‌های زمانی جریان رودخانه‌ها با رژیم جریان طبیعی با عملکرد خوب شبیه‌سازی شد؛ درحالی‌که این عملکرد برای تغییرات دبی در نتیجه ذخیره‌سازی و انحراف آب برای آبیاری به‌ویژه در پایین دست سدها اندکی افت داشت. عملکرد این الگوریتم در پُرکردن سری زمانی روزانه جریان با تغییرات شدید رژیم جریان مانند دبی اوج، مطلوب ارزیابی نشد. این افت عملکرد بیشتر متوجه شرایط هیدرواقلیمی حوزه آبریز مورد مطالعه است تا ساختار الگوریتم. هیدروگراف‌های بازسازی شده امکان تجزیه و تحلیل تغییر و تنوع جریان و برهم‌کنش آن‌ها با متغیرهای آب و هوایی کلیدی را فراهم می‌کنند.

نتیجه‌گیری: الگوریتم جنگل گمشده به‌عنوان یکی از روش‌های بازسازی مبتنی بر یادگیری ماشین دارای اعتبار و عملکرد بالا در بازسازی داده‌های گمشده جریان روزانه رودخانه معرفی شده و می‌توان از آن به‌صورت خودکار و هوشمند در بازسازی نواقص آماری جریان رودخانه در مقیاس روزانه استفاده نمود. پیشنهاد می‌گردد اثرات حوضه‌های مختلف با ویژگی‌های هیدروفیزیکی و اقلیمی خاص در مطالعات آتی بر روی عملکرد الگوریتم جنگل گمشده مورد تجزیه و تحلیل قرار گیرد. بررسی روش پیشنهادی این مطالعه در سایر مناطق هیدرواقلیمی و جغرافیایی، سنجش حساسیت به رژیم بارندگی و جریان رودخانه و در نهایت بررسی عملکرد آن در مقایسه با سایر روش‌های رایج از جمله موارد دیگری است که در مطالعات آتی می‌توان به آن پرداخت.

واژه‌های کلیدی: الگوریتم جنگل تصادفی، جریان رودخانه، داده گمشده، نیکویی برازش، یادگیری ماشین

مقدمه

متغیرهای هیدرومتئورولوژیکی بلندمدت می‌توانند برای درک اقلیم یک منطقه و همچنین برای ارزیابی آسیب‌پذیری منابع آب در منطقه یا جامعه مورد استفاده قرار گیرند (Hamzah *et al.*, 2020). همچنین می‌توان از این داده‌ها برای برنامه‌ریزی و مدیریت منابع آب در سطح حوزه با استفاده از مدل‌های فیزیکی مختلف مانند مدل‌های هیدرولوژیکی و هیدرولیکی استفاده کرد. داده‌های هیدرولیکی به‌عنوان

[DOI: 10.61186/jwmr.15.2.49]

متداول‌ترین داده‌ها راهگشای روابط تجربی تخمین رواناب می‌باشد (Heidari Chenarie *et al*. 2022). با این‌حال، چنین متغیرهایی اغلب با داده‌های از دست رفته مواجه می‌شوند که تجزیه و تحلیل را دشوار یا گاهی اوقات آن را غیرممکن می‌کند (Nadi *et al*., 2022). سری زمانی کامل هیدرولوژیکی برای مدیریت و مدل‌سازی آب، انرژی و سایر منابع طبیعی در یک اقلیم متغیر بسیار حیاتی هستند (Arriagada *et al*., 2019). گمشدگی داده باعث مشکلاتی در تفسیر، واسنجی مدل و آماره‌های اُریب‌دار می‌شود (Dembele *et al*., 2019; Starrett *et al*., 2010). به‌دلایل متعدد از جمله محدودیت منابع اقتصادی و درگیری‌های سیاسی، اداره و مدیریت پراکنده ایستگاه‌های جریان‌سنجی، خاموشی دستگاه‌های اندازه‌گیری، اثرات رویدادهای شدید جوی، دسترسی محدود به دانلود داده‌ها از لاگرهای واقع در مناطق دورافتاده، کمبود ناظران و خطای انسانی، معمولاً داده‌های گمشده در جریان روزانه یافت می‌شوند (Elshorbagy *et al*. 2000; Harvey *et al*., 2012).

به‌طور کلی، مشکل داده‌های از دست رفته در بسیاری از زمینه‌های تحقیقاتی مانند زمینه محیطی (Junninen *et al*., 2004; Plaia and Bondi, 2006; Norazian *et al*, 2008)، آماری (Di Zio *et al*, 2007; Huisman, 2009)، مطالعه پزشکی (Sartori *et al*, 2005; Verboven *et al*, 2007) و صنعتی (Lakshminarayan *et al*, 1999) ظاهر می‌شود. موندلو (Modelo, 2006)، داده‌های گمشده را به‌عنوان یک اصطلاح عمومی تعریف می‌کند. مشاهدات از دست رفته، درک تحلیل داده‌ها را برای تحلیل‌گران دشوار می‌کند (Alibakhshi *et al*., 2019). انواع مشکلاتی که معمولاً با مقادیر از دست رفته همراه هستند عبارتند از: ۱) از دست دادن کارایی ۲) مشکلات در مدیریت و تجزیه و تحلیل داده‌ها ۳) اُریب ناشی از تفاوت بین داده‌های گمشده و کامل (تخمین اُریب‌دار) و ۴) کاهش قدرت آماری یا تخمین ناکارآمد (Hawthorne and Elliott, 2005).

سری زمانی جریان روزانه رودخانه‌های با رژیم جریان طبیعی (Poff *et al*., 1997) برای کاربرد موفقیت‌آمیز روش‌های پُرکردن شکاف مناسب‌تر هستند. در دهه‌های اخیر، جریان رودخانه‌ها در اثر تغییر اقلیم و مداخلات انسانی به‌ویژه در مناطق خشک و نیمه‌خشک به‌طور قابل‌توجهی تغییر یافته است (Kanani *et al*., 2020). رژیم جریان رودخانه، با بزرگی، فرکانس، دوام، زمان‌بندی و نرخ تغییر مشخص می‌شوند که به محرک‌های اقلیم منطقه‌ای و جهانی مانند حالت تغییرپذیری اقلیم، رودبادها، مسیرهای طوفان و رودخانه‌های جوی و همچنین ویژگی‌های حوزه رودخانه مانند کاربری‌ها، زمین‌شناسی، پوشش گیاهی و توپوگرافی پاسخ می‌دهند (Grantham-McGregor *et al*., 2019). در مقابل، در رودخانه‌های با رژیم جریان کنترل شده، تغییر یک یا چند ویژگی رژیم جریان در نتیجه فعالیت‌های انسانی مانند تولید انرژی، حفاظت در برابر سیل، آبیاری، فعالیت‌های صنعتی و تفریحی و شهرنشینی می‌تواند تأثیرات مصنوعی زیادی را ایجاد کند (Mackay *et al*., 2014) و محاسبه خودکار مقادیر جریان گمشده در ایستگاه‌های جریان‌سنجی همجوار را به‌طور قابل‌توجهی پیچیده می‌کند (Harvey *et al*., 2012). در تجزیه و تحلیل سری‌های جریان رودخانه‌ای در مقیاس مکانی بزرگ، هر دو دسته از داده‌های جریان گاهی با همدیگر مخلوط می‌شوند، که از رودخانه‌های با رژیم جریان طبیعی و رژیم جریان کنترل شده حاصل می‌شوند، و روش‌های پُرکردن شکاف‌ها را به چالش می‌کشاند (Dembele *et al*., 2019).

تحقیقات گسترده‌ای در خصوص بازسازی داده‌های هیدرولوژی انجام و هرکدام روش خاصی را برای بازسازی پیشنهاد داده‌اند. تکنیک‌های تکمیل داده‌های جریان گمشده از درون‌یابی ساده تا مدل‌ها و تحلیل‌های آماری پیچیده متفاوت است (Gyau-Boaky and Schultz, 1994). طبقه‌بندی روش‌های موجود برای پرکردن شکاف‌ها در سری‌های زمانی جریان رودخانه با توجه به پیچیدگی آن‌ها توسط (Harvey *et al*., 2012) ارائه شده، که شش دسته از روش‌ها را متمایز کرده است، یعنی استنتاج دستی، تکنیک‌های درون‌یابی پیاپی، فاکتورهای مقیاس‌گذاری، تکنیک‌های هم‌صدک، رگرسیون خطی و مدل‌سازی هیدرولوژیکی. علاوه بر این، تعدادی از روش‌های یادگیری ماشین برای پُرکردن داده جریان گمشده، از جمله شبکه عصبی مصنوعی (به‌عنوان مثال، Aissia *et al*., 2017; Kim *et al*., 2015; Vega-Garcia *et al*., 2019)، مدل‌های جنگل تصادفی (Petty and Dhingra, 2018) و مدل‌های ناپارامتریک تصادفی مانند نمونه‌گیری مستقیم (Dembele *et al*., 2019) استفاده شده‌اند. به‌طور خاص، جنگل تصادفی ارائه شده توسط (Breiman, 2001) یک الگوریتم یادگیری ماشین ناپارامتریک برای شبیه‌سازی داده‌ها براساس ترکیب پیش‌بینی‌های درخت است. این الگوریتم بعدها توسط (Stekhoven and Buhlmann, 2012) به الگوریتم جنگل گمشده برای بازسازی مقدار گمشده در سری داده‌های نوع مختلط گسترش یافت. الگوریتم جنگل گمشده، رویکرد متفاوتی نسبت به جنگل تصادفی با بازنویسی مسئله داده‌های گمشده به‌عنوان یک مسئله پیش‌بینی دارد. داده‌ها با رگرسیون هر متغیر به‌نوبه خود در برابر همه متغیرهای دیگر بازسازی می‌شوند و سپس پیش‌بینی داده‌های گمشده برای متغیر وابسته با استفاده از برازش جنگل گمشده انجام می‌گیرد (Tang and Ishwaran, 2017). مزایای بالقوه مدل‌های جنگل گمشده نسبت به سایر جایگزین‌ها برای پُرکردن داده‌های جریان گمشده در مناطق بزرگ عبارتند از: (۱) آن‌ها می‌توانند به‌سرعت حجم زیادی از داده‌ها را مدیریت کنند و بازسازی داده‌های گمشده بدون نظارت و خودکار است (Sidibe *et al*., 2018)، (۲) آن‌ها می‌توانند شکاف‌های چندگانه داده را مدیریت کنند (Tang and Ishwaran, 2017)، (۳) پیاده‌سازی آن‌ها در زبان‌های محاسباتی مانند R آسان است، زیرا به تنظیم اولیه و واسنجی پارامترها نیاز ندارند (Muñoz *et al*., 2018) و (۴) آن‌ها به عملکرد پیش‌بینی رقابتی دست می‌یابند و از نظر محاسباتی کارآمد هستند و آن‌ها را برای کارهای پیش‌بینی دنیای واقعی مناسب می‌کند (Sidibe *et al*., 2018).

جنگل تصادفی در زمینه‌های علمی مختلف مانند کیفیت هوا (Koçak, 2024) پزشکی (Deshmukh *et al*., 2019; Williams et al., 2023; Stekhoven and Bühlmann,

[DOI: 10.61186/jwmr.15.2.49]

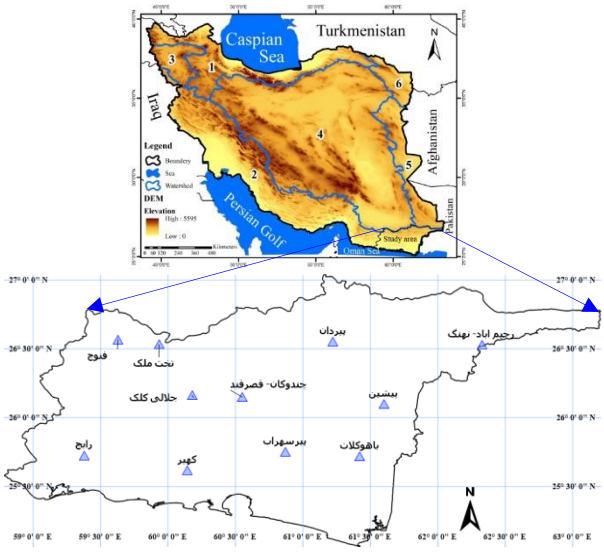
2012; Waljee *et al*., 2013)، حفاظت از اطلاعات حساس (Marino *et al*., 2019) و شیمی غذا (Tao *et al*., 2019) به‌کار گرفته شده است. در زمینه منابع آب (Tyralis *et al*., 2019) جنگل‌های تصادفی اخیراً برای بازسازی جریان‌های ماهانه در مناطق با اقلیم‌های مختلف (Sidibe *et al*., 2018) و برای پیش‌بینی سیل ناگهانی (Munoz *et al*., 2018) آزمایش شده‌اند. روش‌های مدرن، مانند جنگل گمشده، عملکرد بهتری نسبت به روش‌های سنتی قدیمی‌تر نشان داده‌اند (Waljee *et al*., 2013). به‌طور خاص، الگوریتم جنگل گمشده عملکرد بهتری نسبت به‌روش شناخته شده نزدیکترین همسایه (kNN) (Troyanskaya *et al*.,2001)، و الگوریتم پارامتری انتساب چندمتغیره مبتنی بر معادلات زنجیره‌ای (MICE) (Tang and Ishwaran, 2017; Van Buuren, 2007) دارد. در هیدرولوژی، جنگل گمشده عملکرد شبیه به روش‌های مدرن مانند انتساب چندمتغیره مبتنی بر معادلات زنجیره‌ای دارد که توسط (Sidibe *et al*., 2018) برای تکمیل داده‌های جریان ماهانه نشان داده است. پُرکردن شکاف‌ها در سری‌های زمانی جریان روزانه در مناطق با اقلیم‌های مختلف چالش‌برانگیزتر از کاربردهای قبلی است، زیرا تغییرپذیری مکانی و زمانی جریان رودخانه بیشتر است. واضح است که، داده‌های شکاف‌دار هنگام بررسی روندها در مقیاس سالانه دقیق‌تر هستند، سپس در مقیاس ماهانه و در نهایت، مقیاس روزانه، که در آن نتایج کمترین رضایت‌بخشی را دارند (Zhang and Post, 2018).

هدف از پژوهش حاضر، ارزیابی قابلیت اطمینان به الگوریتم جنگل گمشده به‌عنوان یک الگوریتم یادگیری ماشین برای پُرکردن خودکار شکاف در سری زمانی روزانه جریان رودخانه در یک اقلیم خشک و کم داده، با استفاده از جریان سنج‌های موجود در رودخانه‌های با رژیم دائمی و موقتی است.

مواد و روش‌ها

منطقه مورد مطالعه

بر اساس تقسیمات هیدرولوژیکی، تعداد شش حوزه آبریز اصلی در گستره ایران وجود دارد. این حوزه‌های آبریز به‌ترتیب مساحت عبارتند از ۱) حوزه آبریز فلات مرکزی، ۲)حوزه آبریز خلیج فارس و دریای عمان، ۳) حوزه آبریز دریای خزر، ۴) حوزه آبریز مرزی شرق، ۵) حوزه آبریز دریاچه ارومیه و ۶) حوزه آبریز قره قوم. این حوزه‌های آبریز شش‌گانه به حوزه‌های فرعی از درجات مختلف تقسیم‌بندی شده‌اند. حوزه آبریز فلات مرکزی با ۵۱ درصد از گستره ایران، خشک‌ترین حوزه آبریز محسوب می‌شود. حوزه آبریز خلیج فارس و دریای عمان نیز با ۲۶ درصد از گستره ایران، طویل‌ترین حوزه آبریز شناخته می‌شود. منطقه مورد مطالعه این پژوهش (حوزه آبریز بلوچستان جنوبی) از مجموعه حوزه آبریز خلیج فارس و دریای عمان می‌باشد (شکل۱). این حوزه آبریز با کد ۲۹ از تقسیمات فرعی حوزه آبریز خلیج فارس و دریای عمان به‌نام رودخانه‌های بلوچستان جنوبی و با حدود بین سدیج و مرکز پاکستان شناخته می‌شود. مساحت این حوزه آبریز ۴۸۵۲۳/۷ کیلومترمربع و با نام اختصاری رابچ – باهوکلات معروف است. این حوزه آبریز، در گوشه جنوب‌شرقی از حوزه آبریز خلیج فارس و دریای عمان قرار داشته و خشک و کم‌آب می‌باشد. حوزه آبریز خلیج فارس و دریای عمان دارای ۹ حوزه آبریز درجه ۲ می‌باشد که حوزه آبریز رودخانه‌های بلوچستان جنوبی، نهمین حوزه آبریز درجه ۲ از حوزه آبریز خلیج فارس و دریای عمان می‌باشد. حوزه آبریز رودخانه بلوچستان جنوبی از لحاظ هیدروکلیمایی با سایر حوزه‌های آبریز تفاوت دارد. علی‌رغم خشک و کم‌آب بودن این حوزه آبریز، تعدادی رودخانه دائمی و فصلی در این حوزه آبریز جریان دارد که شریان حیاتی سکونت‌گاه‌های نسبتاً پراکنده و خطی در این گستره از سرزمین پهناور ایران می‌باشد. بارش اندک، دما و تبخیر سالانه بالا، نبود جریان‌های دائمی را در این حوزه آبریز سبب می‌گردد. عمده جریان‌های رودخانه‌ای در این حوزه آبریز از نوع فصلی می‌باشد. تنها رودخانه دائمی این منطقه رودخانه سرباز می‌باشد که در نهایت وارد سد پیشین می‌گردد (Damadi *et al*., 2021). به‌طور کلی، تعداد ۱۱ ایستگاه جریان‌سنجی در گستره حوزه آبریز بلوچستان جنوبی وجود دارد که آب‌سنجی را انجام می‌دهند (شکل۱).



شکل ۱– موقعیت منطقه مورد مطالعه و توزیع مکانی ایستگاه‌های جریان‌سنجی
Figure 1. Location of study area and spatial distribution of streamflow gauges

[DOI: 10.61186/jwmr.15.2.49]

روش پژوهش

جنگل‌های تصادفی (RF) درختان تصمیم‌گیری زیادی را رشد می‌دهند و از نتایج آن‌ها میانگین می‌گیرند (Breiman, 2001). هر گره در هر درخت تصمیم‌گیری یک زیرمجموعه تصادفی از متغیرها را انتخاب می‌کند و یک تکنیک بوت استرپ تجمعی (bootstrap aggregation) که یک تکنیک یادگیری گروهی می‌باشد را بر آن‌ها اعمال می‌کنند. الگوریتم جنگل تصادفی (RF) توسط (Stekhoven and Bühlmann, 2012) به الگوریتم جنگل گمشده (MF) برای بازسازی مقدار گمشده در داده‌های مختلط توسعه داده شده است. الگوریتم جنگل گمشده (MF)، شامل آموزش یک جنگل تصادفی به‌صورت تکراری بر روی متغیرهای مشاهده شده برای پیش‌بینی مقادیر از دست رفته است. در این الگوریتم، $X = (X_1 \dots . X_p)$ به‌عنوان مجموعه داده با ابعاد $n \times p$ تعریف می‌گردد که مربوط به p ایستگاه‌های جریان‌سنجی و n جریان‌های روزانه ثبت شده است. برای یک ایستگاه جریان‌سنج (X_s)، عبارت (i_{miss}) مجموعه روزهایی است که ایستگاه S مقادیر گمشده را نشان می‌دهد. سپس مجموعه داده به چهار بخش تقسیم می‌گردد:

۱. $Y_{obs}^{(s)}$: ارزش‌های جریان رودخانه مشاهداتی در ایستگاه جریان‌سنج X_s

۲. $Y_{miss}^{(s)}$: ارزش‌های گمشده جریان رودخانه مشاهداتی در ایستگاه جریان سنج X_s

۳. $X_{obs}^{(s)}$: جریان رودخانه مشاهده شده در ایستگاه جریان‌سنج دیگر در روزهای $\{1. \dots . n\} \backslash i_{miss}(S)$

۴. $X_{miss}^{(s)}$: جریان گمشده در ایستگاه جریان‌سنج دیگر در روزهای $i_{miss}(s)$

قابل‌توجه است که $X_{obs}^{(s)}$ می‌تواند ارزش‌های گمشده داشته باشد و $X_{miss}^{(s)}$ می‌تواند شامل جریان‌های مشاهده شده باشد. هدف ما پُرکردن ارزش‌های از دست رفته $(X_{miss}^{(s)})$ است. برای انجام این‌کار، هدف اصلی این است که یک جنگل تصادفی را برای پیش‌بینی $(Y_{obs}^{(s)})$ از $(X_{obs}^{(s)})$ آموزش داده و سپس از این جنگل تصادفی آموزشی برای پیش‌بینی ارزش‌های گمشده در ایستگاه جریان‌سنج $(X_s(Y_{miss}^{(s)})$ از $(X_{miss}^{(s)})$ استفاده شود. با این وجود ممکن است برخی از مقادیر گمشده در $X_{miss}^{(s)}$ و $X_{obs}^{(s)}$ وجود داشته باشند، که در این‌صورت باید این مقادیر را در مرحله اول به‌صورت زیر پُر شود: میانگین جریان رودخانه روزانه ثبت شده در هر ایستگاه جریان‌سنج X_t در طول دوره مطالعه به هر ارزش گمشده از ایستگاه جریان‌سنج t نسبت داده می‌شود. اکنون، ایستگاه‌های جریان‌سنج با شناسایی آن‌هایی که داده‌های گمشده کمتری دارند، مرتب می‌شوند. برای هر ارزش X_s، ارزش‌های گمشده با برازش یک جنگل تصادفی با وروی $X_{obs}^{(s)}$ و خروجی $X = (X_1 \dots . X_p)$ بازسازی می‌شود. در ادامه، ارزش‌های گمشده $Y_{miss}^{(s)}$ به‌وسیله جنگل تصادفی آموزش دیده با ورودی $X_{miss}^{(s)}$ پیش‌بینی می‌گردد. رویه بازسازی تکرار می‌شود تا زمانی‌که برای اولین بار تفاوت بین داده‌های جدید نسبت داده شده و داده‌های قبلی افزایش یابد. به‌طور دقیق‌تر، فرض کنیم X_k^{imp} داده‌های بازسازی شده قبلی در تکرار k^{th} باشد و X_{k+1}^{imp} بازسازی به‌روزشده در تکرار $k+1$ باشد. تفاوت (Δ) به‌صورت زیر محاسبه می‌گردد:

$$\Delta_k = \frac{\sum i \in X(X_{k+1}^{imp} - X_k^{imp})^2}{\sum i \in X(X_{k+1}^{imp})} \qquad (۱)$$

معیار توقف به‌محض این‌که Δ_{k+1} بزرگ‌تر از Δ_k باشد، برآورده می‌شود. هزار درخت رگرسیونی در تمام محاسبات براساس تجربیات قبلی توسط (Arriagada *et al*., 2019) استفاده شده است و بیشینه تعداد تکرار روی صد تنظیم شده است. زیرا این اعداد به‌اندازه کافی برای اطمینان از تحقق معیار هم‌گرایی در معادله قبل، بزرگ است. این الگوریتم با استفاده از پکیج missForest نسخه 1.5 در محیط نرم‌افزار R پیاده‌سازی شده است.

اغلب، عملکرد الگوریتم‌های بازسازی داده‌های گمشده با سناریوهای داده گمشده مصنوعی به چالش کشیده می‌شوند. جریان‌های روزانه گمشده در منطقه مورد مطالعه در طول سال به‌طور یکنواخت توزیع شده‌اند. دو نوع شکاف مصنوعی تولید شده است. (۱) بخش‌های پیوسته حذف شده: در هر ایستگاه جریان‌سنج فقط یک بخش (با طول‌های ۷، ۱۴، ۲۱، ۳۰، ۶۰ و ۳۶۵ روز) به‌طور تصادفی از کل دوره ثبت شده (۱۹۷۲–۲۰۱۸) حذف شده است. (۲) نقاط داده تکی حذف شده: مقادیر مشاهده شده مربوط به (روزهای ۳۰، ۶۰، ۹۰، ۱۲۰، ۱۸۰ و ۳۶۵) به‌طور تصادفی از کل دوره ثبت شده (۱۹۷۲–۲۰۱۸) در هر ایستگاه جریان‌سنج حذف شدند. الگوریتم MissForest برای پُرکردن شکاف‌های موجود در رکوردها همراه با شکاف‌های مصنوعی استفاده شده است. عملکرد الگوریتم MissForest در پُرکردن داده‌های جریان روزانه با مقایسه سری‌های پُرشده با داده‌های مشاهده شده از طریق آزمون‌های نیکویی برازش (GoF) تست شده است که شامل ضریب تعیین (R^2)، درصد اُریب (PBIAS) و معیار کلینگ– کوپتا (KGE) می‌باشند (Kling *et al*., 2012).

$$R^2 = \left[\frac{\sum_{i=1}^{n}(O_i - \mu_0)(S_i - \mu_s)}{\sqrt{\sum_{i=1}^{n}(O_i - \mu_0)^2}\sqrt{\sum_{i=1}^{n}(O_s - \mu_s)^2}}\right] \qquad (۲)$$

$$PBIAS = \left[\frac{\sum_{i=1}^{n} S_i - O_i}{\sqrt{\sum_{i=1}^{n} O_i}}\right] \times 100 \qquad (۳)$$

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

$$\beta = \frac{\mu_s}{\mu_0} \quad and \quad \gamma = \frac{\sigma_s/\mu_s}{\sigma_0/\mu_0} \qquad (۵)$$

در این‌جا، O و S به‌ترتیب داده‌های مشاهده شده و شبیه‌سازی شده، μ و σ نیز به ترتیب میانگین انحراف معیار و r ضریب همبستگی بین داده‌های مشاهده شده و شبیه‌سازی شده است. β نسبت اُریب و سرانجام γ نسبت تغییرپذیری است. ارزش بهینه از R^2 و KGE برابر یک است، در حالی‌که ارزش بهینه از $PBIAS$ برابر صفر است. ارزش‌های آستانه برای عملکرد رضایت‌بخش، خوب و خیلی‌خوب عبارتند از $0.75 < R^2 \leq 0.85$، $0.60 < R^2 \leq 0.75$ و $R^2 < 0.85$ و $\pm 10 > PBIAS \geq \pm 5$، $\pm 15 > PBIAS \geq \pm 10$ و $PBIAS < \pm 5$ می‌باشند (Moriasi *et al*., 2007). همچنین دو کلاس برای عملکرد بر طبق معیار کلینگ– کوپتا

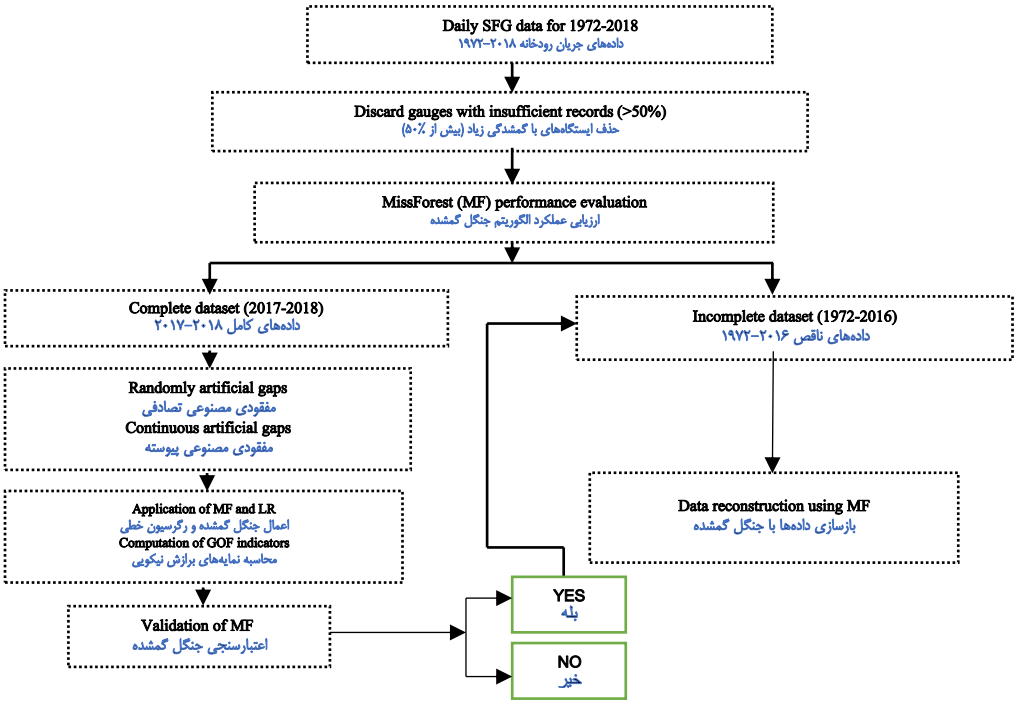
[DOI: 10.61186/jwmr.15.2.49]

(KGE) متمایز شـده اسـت که معیار $KGE > -0.41$ برای عملکرد خوب و$KGE < -0.41$ برای عملکرد بد تعریف شـده است (Knoben *et al.*, 2019). فلوچارت گام‌به‌گام روش کار با الگوریتم جنگل گمشده در شکل (۲) نمایش داده شده است.

نتایج و بحث

درصد گمشدگی جریان رودخانه

درصد گمشدگی داده‌ها، معیار مهم تصمیم‌گیری درباره امکان یا عدم امکان برآورد داده‌های گمشده و همچنین اتخاذ روش‌ها و الگوریتم‌های بهینه جهت برآورد و تخمین آن‌ها می‌باشد. ادبیات پژوهشی درباره درصد گمشدگی داده‌ها و به‌عبارتی ارزش‌های گمشده، متنوع است. شفر (Schafer, 1997) بیان داشته است که فقدان ۵ درصد یا مقدار کمتر از آن در سری‌های زمانی، ناچیز بوده و عملاً خللی در نتایج ایجاد نمی‌کند. بنت (Bennett, 2001) اظهار کرده است که هرگاه درصد داده‌های گمشده در یک سری زمانی بیش از ۱۰ درصد باشد، انتظار می‌رود تحلیل آماری آن سری زمانی مغرضانه (اُریب‌دار) باشد. دانگ و پنگ (Dong and Peng, 2013) توافق نموده‌اند که گمشدگی داده‌ها به‌میزان ۲۰ درصد یک امر رایج در تحقیقات است؛ درحالی‌که ویدامن (Widaman, 2006) داده‌های گمشده را بر اساس درصد گمشدگی در یک جدول بیان کرده است. یک قانون سرانگشتی بیان می‌دارد که هنگامی‌که داده‌های یک سری زمانی شامل کمتر از ۵ درصد گمشدگی باشد که به مقادیر مشاهده شده و مشاهده نشده بستگی ندارد، تحلیل موردی کامل نیز ممکن است یک رویکرد قابل قبول باشد (Graham, 2009). بنابراین، چنانچه داده‌های گمشده کمتر از ۵ درصد حجم یک سری زمانی را شامل شوند، نیازی به بازسازی و برآورد آن‌ها وجود ندارد؛ مگر اینکه ارزش‌های گمشده بسیار حیاتی و تعیین‌کننده باشند. از سوی دیگر، چنانچه داده‌های گمشده بیش از ۲۵ درصد حجم یک سری زمانی باشد، امکان بازسازی و برآورد آن‌ها سخت و پیچیده بوده و قابلیت اطمینان به داده‌های برآورد شده کاهش می‌یابد. بررسی و تحلیل گمشدگی داده‌های جریان رودخانه در ایستگاه‌های هیدرومتری حوزه بلوچستان جنوبی به‌شرح جدول (۲) می‌باشد. بر اساس قواعد سرانگشتی موجود، چنانچه بیش از ۵۰٪ داده‌ها گمشده باشند نبایستی از بازسازی استفاده نمود. بر این اساس، ایستگاه‌های ردیف ۸ تا ۱۱ از فرایند بازسازی کنار گذاشته می‌شوند. میزان درصد گمشدگی در داده‌های ایستگاه‌های ردیف ۸ تا ۱۱ متناسب با دوره آماری آن‌ها بسیار اندک است. منتها از سوابق آماری طولانی برخودار نیستند. این ایستگاه‌ها عمدتاً در سال‌های بعد از ۲۰۰۰ میلادی تأسیس شده‌اند. از آن‌جا که درصد گمشدگی داده‌های جریان در دهه‌های اخیر، بسیار پایین می‌باشد بنابراین ضرورتی بر دخالت آن‌ها در فرایند مدل‌سازی وجود ندارد.



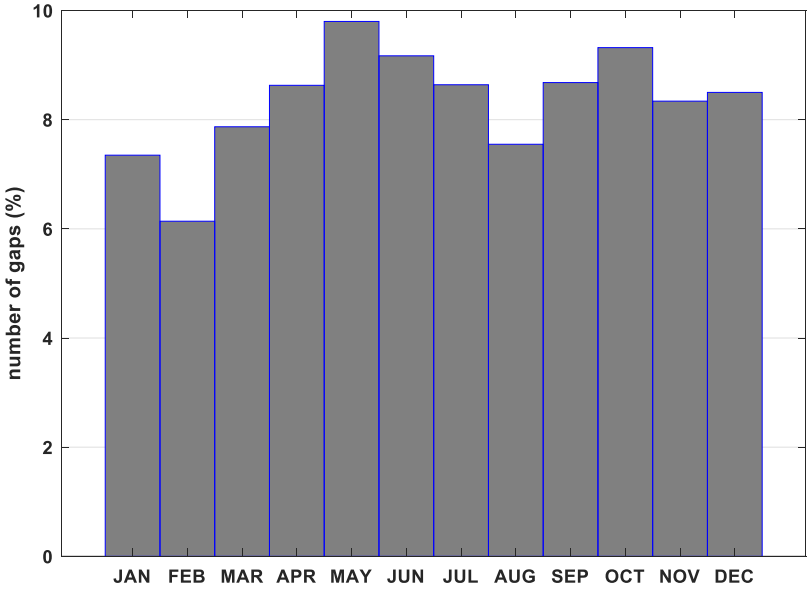
شکل ۲- نمودار گردشی مراحل انجام کار با الگوریتم جنگل گمشده
Figure 2. Flowchart illustrating the workflow step by step of MissForest Algorithm

[DOI: 10.61186/jwmr.15.2.49]

جدول ۲- درصد گمشدگی در داده‌های روزانه جریان ایستگاه‌های هیدرومتری حوزه آبریز بلوچستان جنوبی
Table 2. Missingness percentage in daily streamflow data of gauges in study area

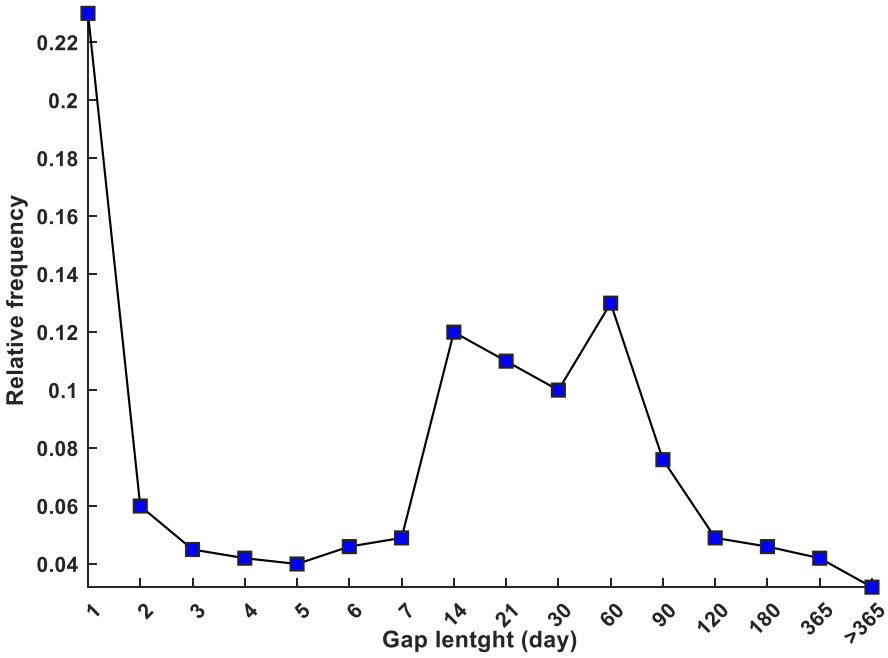
ردیف No.	ایستگاه gauge	تأسیس (سال) Established (Year)	گمشدگی (سال) Missingness (Year)	درصد گمشدگی Missingness (%)	تأیید (بله/خیر) Verification (Y/N)
1	باهوکلات BahoKalat	1972	8	49.64	بله (Y)
2	پیردان Pirdan	1975	5	17.04	بله (Y)
3	پیشین Pisshin	1974	9	46.52	بله (Y)
4	پیرسهراب PirSohrab	1981	0	47.34	بله (Y)
5	چندوکان ChanDokan	1972	12	41.09	بله (Y)
6	کهیر Kahir	1973	15	47.48	بله (Y)
7	کاریانی Karyani	1982	14	49.35	بله (Y)
8	تخت‌ملک Takht Malek	1998	1	61.80	خیر (N)
9	رحیم‌آباد RahimAbad	2004	5	82.60	خیر (N)
10	فنوج Fanoj	2008	0	78.25	خیر (N)
11	جلالی‌کلک JalaiKalak	2005	2	83.52	خیر (N)

توزیع گمشدگی آماری در طول ماهای سال در شکل (۳) نشان داده شده است. جریان‌های روزانه گمشده در منطقه مورد مطالعه در طول سال به‌طور یکنواخت توزیع شده‌اند. تفاوت معنی‌دار بین درصد گمشدگی در ماه‌های سال وجود ندارد. دامنه گمشدگی بین ۶ تا کمتر از ۱۰ درصد متغیر است. کمترین و بیشترین درصد گمشدگی به‌ترتیب مربوط به ماه‌های فصل زمستان و بهار می‌باشد. به‌طور کلی ماه‌های فصل زمستان و تابستان نسبت به ماه‌های فصل بهار و پاییز، درصد گمشدگی پایین‌تری دارند. این مشخصه به ویژگی‌های هیدرواقلیمی منحصربه‌فرد حوزه آبریز بلوچستان جنوبی مربوط می‌شود. رژیم بارش در این حوزه آبریز غلبه بر تمرکز در فصل زمستان و تابستان دارد. همان‌گونه که بیان شد، با افزایش طول دوره آماری میزان درصد گمشدگی نیز افزایش می‌یابد. علاوه بر طول دوره آماری، معیار حداقل گمشدگی آماری نیز در موفقیت بازسازی داده‌های گمشده حائز اهمیت است. فراوانی نسبی گروه‌های مشخص از طول گمشدگی آماری در شکل (۴) برای ایستگاه‌های انتخاب شده (۷ ایستگاه باقیمانده) استخراج شده است. در بین گمشدگی کمتر از ۷ روز، بیشترین فراوانی مربوط به گمشدگی یک‌روزه می‌باشد. انتظار می‌رود، این ویژگی ناپیوستگی زمانی گمشدگی، موفقیت روش بازسازی را تحت تأثیر قرار دهد. گمشدگی آماری با طول کمتر از ۱۴، ۲۱، ۳۰، ۶۰ و ۹۰ روزه سهم بالاتری از کل گمشدگی را به‌خود اختصاص می‌دهند.



شکل ۳- توزیع داده‌های گمشده در طول ماه‌های سال در ایستگاه‌های جریان‌سنجی منطقه مورد مطالعه
Figure 3. Distribution of missed data in streamflow gauges along the months of the year in study area

[DOI: 10.61186/jwmr.15.2.49]



شکل ۴- فراوانی نسبی طول گمشدگی آماری در ایستگاه‌های جریان‌سنجی منطقه مورد مطالعه
Figure 4. Relative frequency of gap lengths in streamflow gauges in study area

عملکرد الگوریتم جنگل گمشده

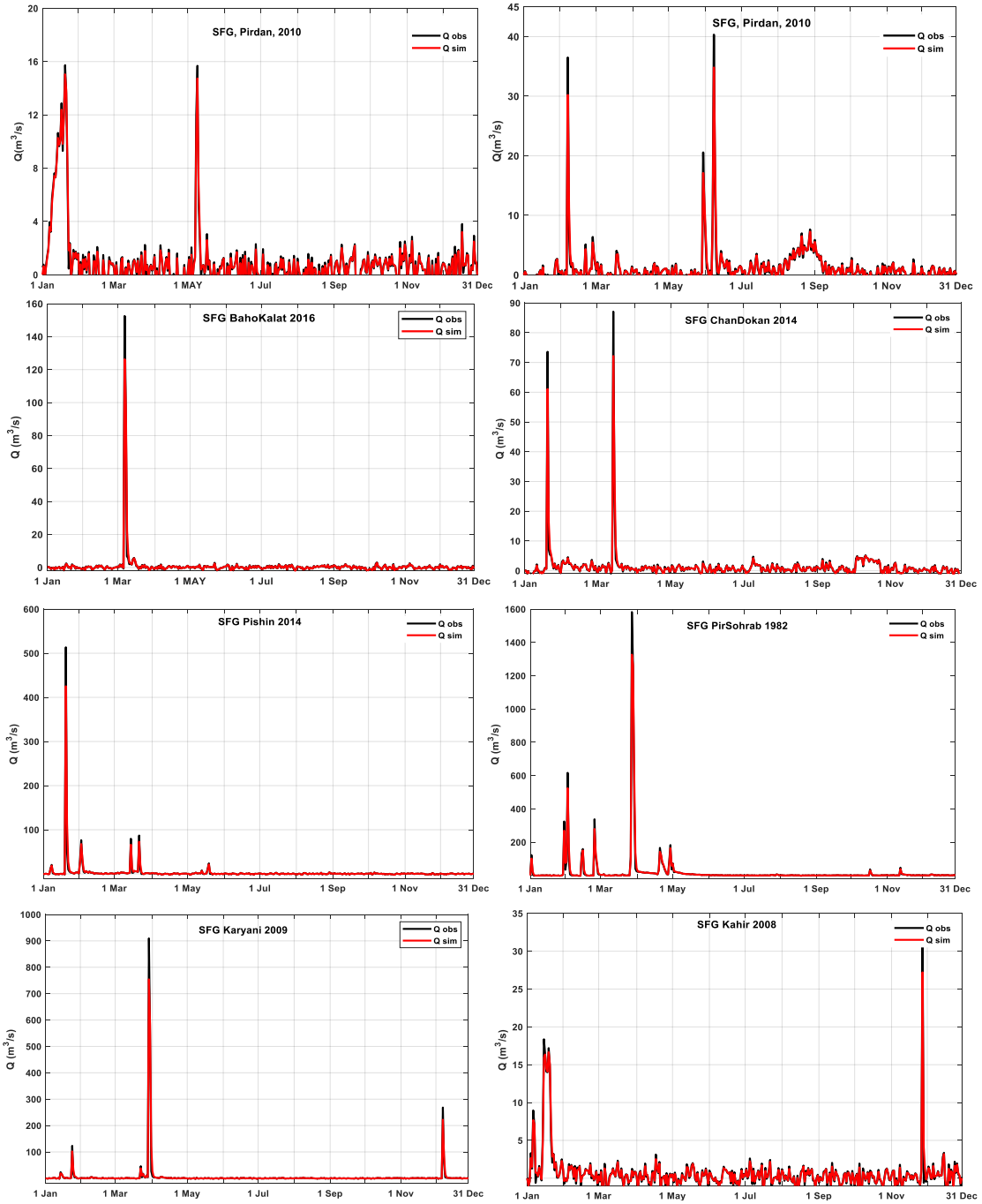
نتایج تجزیه و تحلیل درصد گمشدگی داده‌ها، در اتخاذ شیوه ارزیابی عملکرد الگوریتم جنگل گمشده تأثیرگذار است. همان‌گونه که در شکل (۲) مشخص شده است، عملکرد الگوریتم جنگل گمشده در بازسازی داده‌های گمشده با سناریوهای داده گمشده مصنوعی به چالش کشیده شده است. چنانچه الگوی گمشدگی داده‌ها تابعی از رژیم‌های متفاوت جریان رودخانه‌ها باشد لازم است شرایط اعتبارسنجی برای رژیم‌های آبدهی متفاوت، مجزا لحاظ گردد. در این پژوهش، الگوی گمشدگی زمانی (ماهانه) داده‌ها یکدست بود و تابعی از رژیم‌های بارش و جریان رودخانه نمی‌باشد. بنابراین، تنها معیار طول گپ‌های آماری تعیین کننده سناریوهای اعتبارسنجی می‌باشد. بنابراین، دو نوع شکاف مصنوعی در داده‌ها تولید شده است. ۱) حذف بخش‌های پیوسته: در هر ایستگاه جریان‌سنج فقط یک بخش (با طول‌های ۷، ۱۴، ۲۱، ۳۰، ۶۰ و ۳۶۵ روز) به‌طور تصادفی از کل دوره ثبت شده (۲۰۱۸-۱۹۷۲) حذف شده است. ۲) حذف تکی نقاط داده: مقادیر مشاهده شده مربوط به (روزهای ۳۰، ۶۰، ۹۰، ۱۲۰، ۱۸۰ و ۳۶۵) به‌طور تصادفی از کل دوره ثبت شده (۲۰۱۸-۱۹۷۲) در هر ایستگاه جریان‌سنج حذف شده‌اند. در نهایت، الگوریتم MissForest برای پُرکردن شکاف‌های مصنوعی موجود در رکوردها استفاده شده است. نتایج به‌دست آمده از طریق این الگوریتم با مقایسه نتایج محاسبه شده از طریق رگرسیون خطی (LR) در روش log-log space محک زده شده است. رگرسیون خطی با استفاده از تمام روزهای غیرگمشده تصادفی در ایستگاه آب‌سنجی موردنظر و نزدیکترین ایستگاه آب‌سنجی محاسبه شده است. نزدیکترین ایستگاه اندازه‌گیری ایستگاهی بود که کمترین فاصله خطی را تا گیج آب‌سنجی موردنظر داشت.

برای بررسی این‌که آیا الگوریتم MissForest به تعداد روزهای گمشده در درون پرونده‌ای که در حال پُرشدن است حساس است یا خیر؛ تعیین شده است که یک رکورد چقدر باید باشد (چند روز گمشده) قبل از این‌که MissForest بتواند برای تکمیل داده‌های گمشده باقی‌مانده استفاده شود، با نگه‌داشتن (برای آموزش) تعداد متفاوت سال غیرگمشده (۲، ۳، ۴، ۱۰، ۱۵، ۲۰، ۳۰ و ۴۷ سال) درحالی‌که تعداد روزهای گمشده شبیه‌سازی شده (۳۰، ۱۸۰ و ۳۶۵ روز) را پیش‌بینی می‌کند. همچنین دقت پیش‌بینی جریان‌های گمشده مربوط به تعداد رکوردهای گنجانده شده در MissForest از طریق افزایش تدریجی تعداد ایستگاه‌های آب‌سنجی مورد استفاده به‌عنوان پیش‌بینی کننده (۱ تا ۶ ایستگاه آب‌سنجی) افزایش داده شده است. باز هم عملکرد MissForest از مجموعه ۷ ایستگاه آب‌سنجی توسط مقایسه سری‌های پُرشده و مشاهده شده با استفاده از شاخص‌های نیکویی بزارش شامل R2، PBIAS و KGE آزمایش شده است.

شکل (۵) هیدروگراف‌های مشاهده شده و شبیه‌سازی شده در ایستگاه‌های جریان‌سنج را در بخش‌های مختلف حوزه آبریز بلوچستان جنوبی نشان می‌دهد. ارزیابی عملکرد MissForest در پُرکردن شکاف‌های سری زمانی جریان روزانه از این طریق بررسی شده است. به‌طور کلی، شکل هیدروگراف‌های مشاهده شده به‌خوبی با زمان‌بندی خوب و نمایش فصلی سالانه در همه موارد بازتولید شده و نشان می‌دهد که ریز واحدهای هیدرولوژیکی که جریان‌سنج‌ها در آن‌ها واقع شده‌اند، کنترل‌های مهمی بر عملکرد MissForest

[DOI: 10.61186/jwmr.15.2.49]

ندارند. هیدروگراف‌های شبیه‌سازی شده، مطابقت کامل با جریان‌های کم را در رژیم‌های موقتی تحت سلطه بارندگی کم و حوزه آبریز خُرد و همچنین جریان‌های زیاد در حوزه آبریزهای بزرگ‌تر و پربارش‌تر داشته‌اند.



شکل ۵- هیدروگراف‌های اصلی و شبیه‌سازی شده ایستگاه‌های آب‌سنجی در بازه‌های زمانی مختلف
Figure 5. Observed and simulated hydrographs in streamflow gauges during different time periods

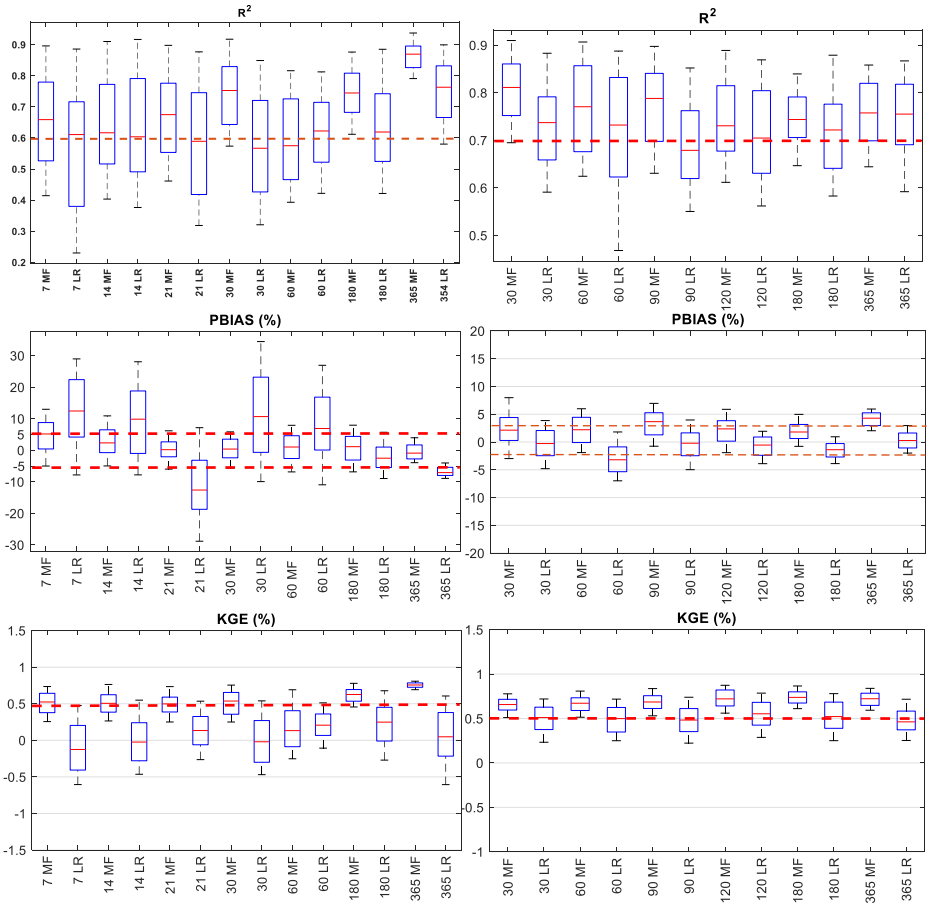
شکل (۶) عملکرد الگوریتم جنگل گمشده (MF) و رگرسیون خطی (LR) را در فضای log-log برای پُرکردن بخش‌های پیوسته حذف شده و نقاط داده منفرد نشان می‌دهد. نتایج نشان می‌دهد که الگوریتم جنگل گمشده، عملکرد خوبی (ارزش‌های متوسط برای $R^2 > 0.6$، $PBIAS < \pm 5$ و $KGE > 0.5$) در پُرکردن نقاط داده منفرد (تکی) و بخش‌های پیوسته داشته است. این نتایج نشان می‌دهد که عملکرد MissForest نسبت به طول و مقدار داده‌های از دست رفته خیلی حساس نیست. الگوریتم MissForest در فضای log-log بهتر از رگرسیون خطی عمل کرده است. به‌طور متوسط، همه شاخص‌های عملکرد بهتر و پراکندگی کوچک‌تر بود. در مقایسه با رگرسیون خطی در فضای log-log، الگوریتم

[DOI: 10.61186/jwmr.15.2.49]

MissForest عملکردهای حداقل و حداکثر بالاتری را تولید کرد. با این‌حال، تجزیه و تحلیل ما نشان می‌دهد که رگرسیون خطی نتایج نسبتاً خوبی برای پُرکردن شکاف‌های یک‌روزه ایجاد می‌کند. این بدان معنی است که از داده‌های یک ایستگاه آب‌سنجی همجوار می‌توان برای پُرکردن شکاف‌ها در این موارد استفاده نمود. علاوه بر این، MissForest تمایل دارد شکاف‌های پُرکردن یک روزه را بیش از حد تخمین بزند. برای الگوریتم MissForest، همبستگی زمانی مهمتر است.

شکل (۷) عملکرد MissForest را برای تعداد متفاوت روزهای غیر گمشده نشان می‌دهد. در مورد رکوردهای کوتاه، یعنی ۲ تا ۵ سال طول، عملکرد MissForest با تعداد روزهای غیرگمشده برای بازسازی داده‌های گمشده به‌طور قابل‌توجهی افزایش داشت. با این‌حال، زمانی‌که رکوردها طولان‌تر از ۱۵ سال می‌شوند، PBIAS و KGE تمایل دارند به یک مقدار ثابت در محدوده رضایت‌بخش یا بهتر همگرا شوند. نتایج KGE نشان می‌دهد که اغلب بیش از ۱۵ سال روزهای غیر از دست رفته برای تولید عملکرد کافی مورد نیاز است.

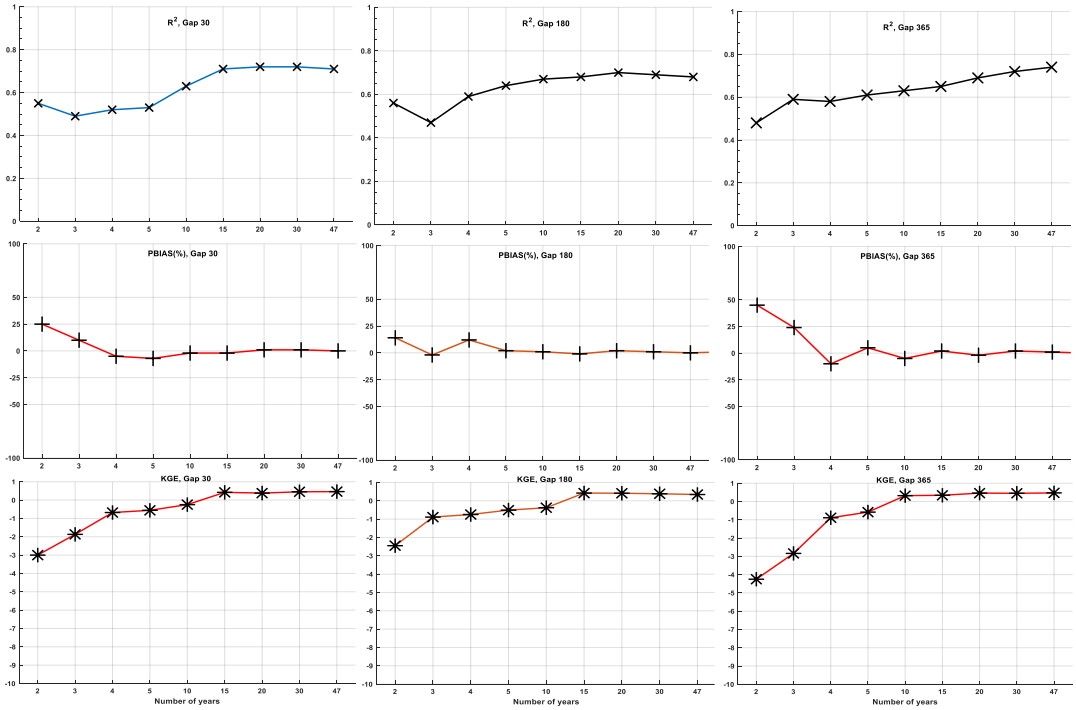
شکل (۸) نشان دهنده R^2، $PBIAS$ و KGE از الگوریتم MissForrest با استفاده از تعداد متفاوت از رکوردها به‌عنوان پیش‌بینی کننده برای پُرکردن شکاف‌ها از ۲٪ (۳۶۵)، ۵٪(۸۵۸) و ۵۰٪(۸۵۸۴) از داده‌ها می‌باشد که به‌طور تصادفی در ایستگاه آب‌سنجی معین (پیردان) برداشت شده است. عملکرد MissForest با تعداد گیج‌های پیش‌بینی افزایش می‌یابد. به‌عنوان مثال، هنگامی‌که تعداد گیج‌های بیشتری به‌عنوان پیش‌بینی کننده استفاده می‌شود، شاخص‌های عملکرد MissForest تمایل دارند تا به یک مقدار ثابت همگرا شوند.



شکل ۶- عملکرد الگوریتم جنگل گمشده (MF) و رگرسیون خطی (LR) برای بخش‌های پیوسته حذف شده (چپ) و تکی حذف شده (راست)

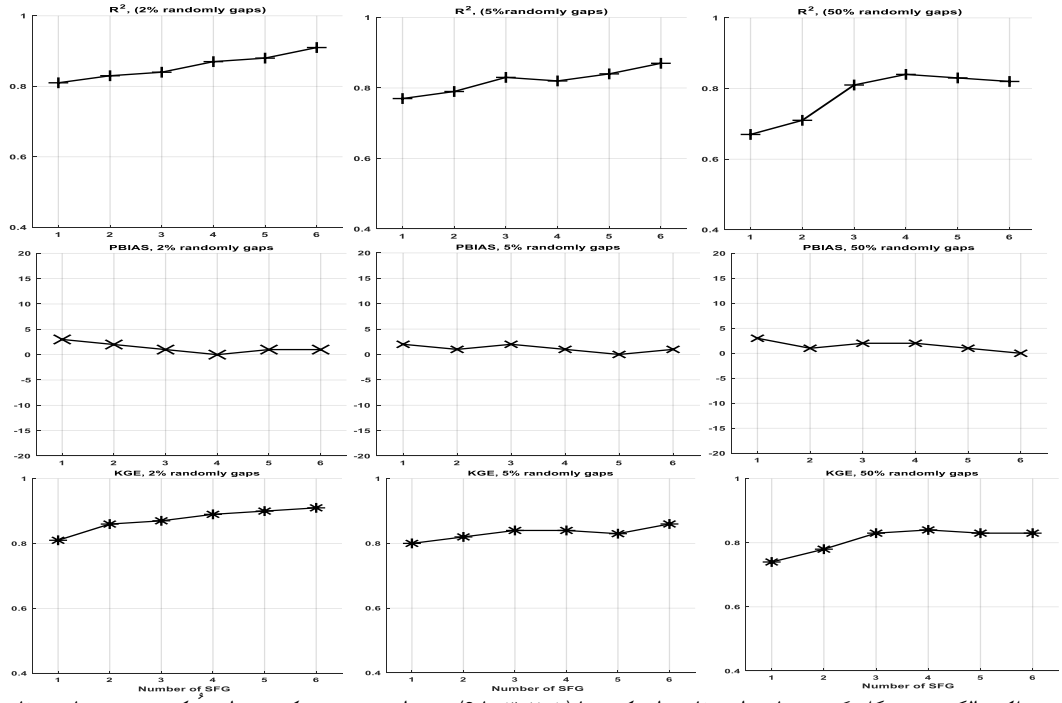
Figure 6. Performance of MissForest (MF) and Linear Regression in log–log space (LR) for removed contiguous segments (left) and removed single data points (right)

[DOI: 10.61186/jwmr.15.2.49]



شکل ۷- عملکرد جنگل گمشده با تعداد مختلف از سال غیرگمشده (۲، ۳، ۴، ۵ و ۴۷ سال) برای تعداد مختلف از روزهای گمشده شبیه‌سازی شده ۳۰ روزه (سمت چپ)، ۱۸۰ روزه (وسط) و ۳۶۵ روزه (سمت راست).

Figure 7. MissForest performance for different numbers of non-missing years (2, 3, 4, 5, and 47 years) whilst predicting on different numbers of simulated missing days 30 (Left), 180 (center) and 365 (right).



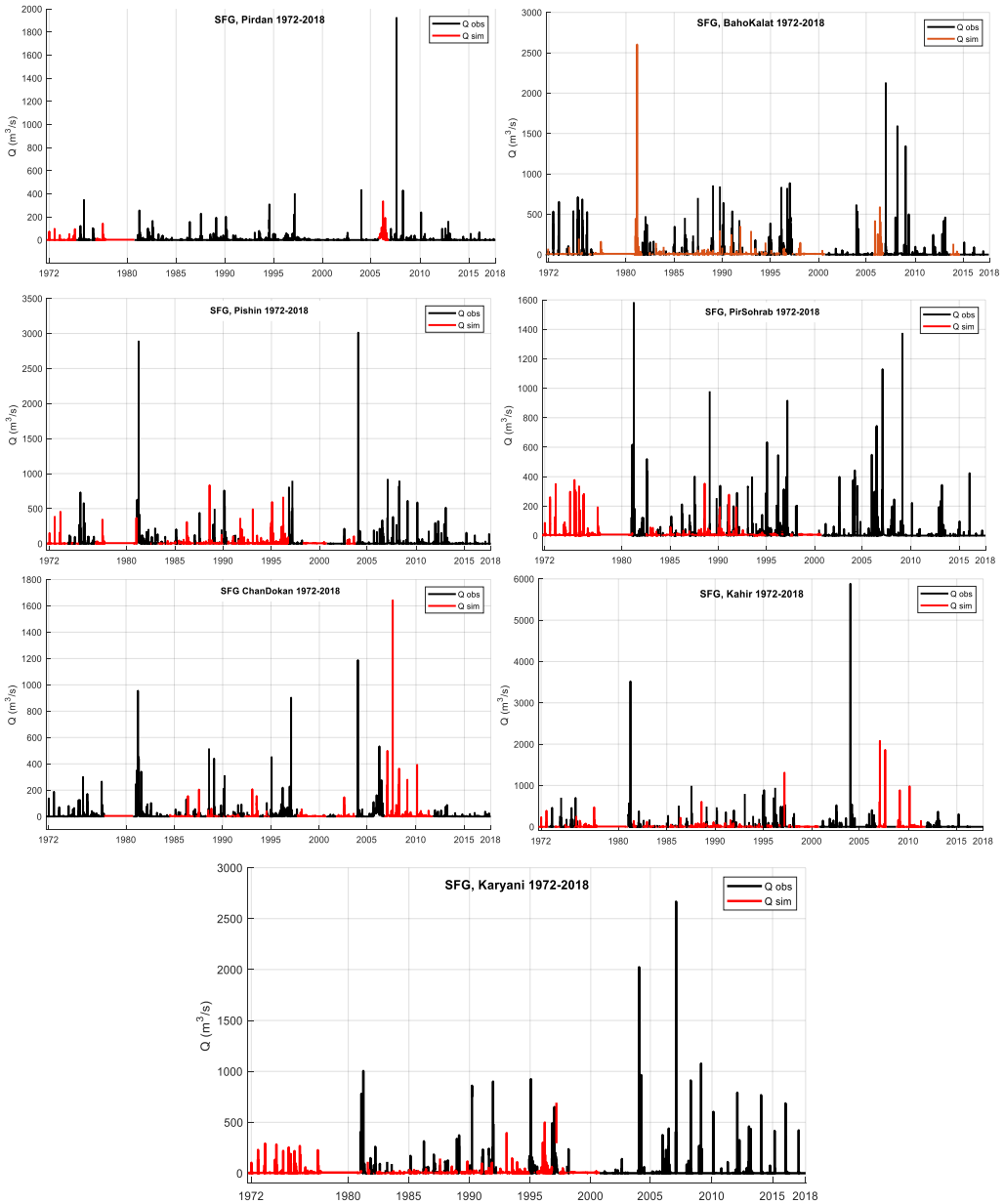
شکل ۸- عملکرد الگوریتم جنگل گمشده با تعداد متفاوت از رکوردها (۱، ۲، ۳، تا ۶) به‌عنوان پیش‌بینی کننده برای پرکردن درصدهای متفاوت از گپ داده‌ها که به‌طور تصادفی انتخاب شده‌اند.

Figure 8. MissForest performance for different number of records (1, 2, 3 to 6) as predictors for filling gaps of different percentage of data randomly taken out at a given gauge

[DOI: 10.61186/jwmr.15.2.49]

بازسازی رکوردهای گمشده جریان رودخانه

سری‌های زمانی کامل از جریان روزانه رودخانه برای مدیریت آب، انرژی و منابع طبیعی بسیار مهم می‌باشند. از آن‌جا که الگوریتم جنگل گمشده به‌طور کلی عملکرد رضایت‌بخش تا خوب در پرکردن شکاف‌های مصنوعی سری‌های زمانی جریان روزانه ارائه کرد، رکوردهای موجود در ۷ ایستگاه جریان‌سنج با درصد گمشدگی کمتر از ۵۰٪ درصد واقع در حوزه آبریز بلوچستان جنوبی مورد بازسازی قرار گرفته‌اند. شکل (۹) هیدروگراف‌های مشاهده شده و بازسازی شده را در طول دوره ۱۹۷۲–۲۰۱۸ در ایستگاه‌های آب‌سنجی هفت‌گانه حوزه بلوچستان جنوبی نشان می‌دهد. در بین ایستگاه‌های انتخابی برای بازسازی داده‌های گمشده، درصد گمشدگی داده‌های روزانه در ایستگاه پیردان واقع بر روی رودخانه دائمی سرباز در حداقل بود. سایر ایستگاه‌ها در دامنه ۴۰٪ تا ۵۰٪ درصد گمشدگی تفاوت معنی‌داری با همدیگر نداشتند. ماهیت جریان دائمی تنها در ایستگاه آب‌سنجی پیردان یافت شد. در سایر ایستگاه‌های آب‌سنجی جریان‌ها عمدتاً رژیم موقتی دارند. برخلاف تصور قبلی، عملکرد بازسازی داده‌های گمشده در رژیم‌های موقتی نسبت به رژیم‌های دائمی با چالش‌های کمتری مواجه بود.



شکل ۹– هیدروگراف مشاهده شده و بازسازی شده در طول دوره ۱۹۷۲–۲۰۱۸ در ایستگاه‌های آب‌سنجی حوزه مورد مطالعه

Figure 9. Observed and reconstructed hydrographs over the 1972–2018 study period at the gauge stations of studied watershed

[DOI: 10.61186/jwmr.15.2.49]

نتیجه‌گیری کلی

محققان اغلب آستانه‌ای را برای درصد قابل قبول از داده‌های گمشده جهت در نظر گرفتن یک ایستگاه جریان‌سنج قابل استفاده، لحاظ می‌کنند. به‌عنوان مثال، آستانه ۱٪ (Petrone *et al*., 2010)، ۵٪ (Ukkola *et al*., 2016)، ۱۰٪ (Déry *et al*., 2009)، ۱۵٪ (Liu and Zhang, 2017) و ۲۰٪ (Lopes *et al*., 2016) در مطالعات قبلی اتخاذ شده است. در این مطالعه ارائه شده، آستانه ۵۰٪ را اتخاذ شده است که به ما امکان می‌دهد با ۷ ایستگاه جریان‌سنج در حوزه آبریز بلوچستان جنوبی کار کنیم. تحت چنین شرایطی، ما نشان دادیم که منطقه مورد مطالعه یک منطقه کم‌داده با دسترسی ضعیف به سوابق جریان روزانه است، که بسیار کمتر از استانداردهای موردنظر توصیه شده توسط سازمان جهانی هواشناسی است. تراکم گیج بسیار پایین و موجودیت داده یک سناریوی چالش برانگیز برای روش‌های پُرکردن شکاف‌های موجود در سری‌های زمانی است. از آن‌جا که منطقه مورد مطالعه فاقد اطلاعات اولیه برای طبقه‌بندی مناسب از جریان‌ها به دسته‌های طبیعی و غیرطبیعی بود، هدف این مطالعه به کارگیری الگوریتم MissForest برای پُرکردن تمام شکاف‌ها با استفاده از تمام گیج‌های مورد با اطلاعات کافی، یعنی درصد داده گمشده کمتر از ۵۰٪ بوده است.

داده‌های از دست‌رفته به‌طور تصادفی در سری‌های زمانی جریان رودخانه رخ داده‌اند و بنابراین فرض بر این بود که آن‌ها به‌طور یکنواخت توزیع شده‌اند. با این‌حال، ممکن است تحت شرایطی خاص (مثلاً در حین سیل و بعد از سیل به‌دلیل اختلال فیزیکی در تجهیزات) اتفاق بیفتد که داده‌های از دست رفته توزیعی متفاوت با یکنواخت را ارائه دهند. در چنین مواردی، انتظار می‌رود که پیش‌بینی روزهای گمشده واقعی به‌طور متوسط بیشتر از روزهای غیرمفقود واقعی باشد. در این پژوهش از روش پیشنهادی (Stekhoven and Bulmann, 2012) پیروی شده است و در نتیجه، هیچ تغییری در مجموعه داده‌ها اعمال نشده است. با این‌حال، MissForest می‌تواند تخمین‌های رگرسیونی بسیار اُریب و پوشش‌های فاصله‌ای اطمینان با اُریب رو به پایین برای متغیرهای دارای چولگی بسیار در مدل‌های غیرخطی تولید کند (Hong and Lynn, 2020). بنابراین، این روش بایستی با احتیاط اعمال شود. به‌طور خاص، داده‌ها نبایستی خیلی چولگی داشته باشند و شکاف داده‌ها نبایستی توزیع اُریب‌دار داشته باشد. از این نظر، انتظار نمی‌رود که تغییر جریان روزانه باعث ایجاد توزیع‌های با چولگی زیاد شود (Blum *et al*., 2017).

از این‌رو، الگوریتم جنگل گمشده به‌عنوان یک الگوریتم یادگیری ماشین تصادفی ناپارامتریک، برای پُرکردن شکاف‌ها در سری‌های زمانی جریان روزانه در هفت ایستگاه آب‌سنجی حوزه بلوچستان جنوبی در دوره ۲۰۱۸–۱۹۷۲ اعمال و عملکرد آن ارزیابی شد. بازسازی جریان‌های دائمی (ایستگاه پیردان) چالش برانگیزتر نسبت به جریان‌های موقتی بود. عملکرد الگوریتم جنگل گمشده در پُرکردن شکاف داده به‌شدت تابع همبستگی زمانی بوده و از این‌رو در بازسازی شکاف‌های یک روزه چندان موفق نبوده و تخمین‌های بیش از مقدار واقعی را ارائه داد. بنابراین الگوریتم MissForest با توجه به ضریب شاخص‌های نیکویی برازش که شامل ضریب تعیین (R^2)، درصد اُریب (PBIAS) و معیار کلینگ– کوپتا (KGE) می‌باشد به عملکرد رضایت‌بخش تا خوب دست یافت. به‌طور قابل‌توجهی، عملکرد MissForest در پُرکردن جریان‌های روزانه در منطقه کم‌داده– در این مورد بلوچستان جنوبی– با روش‌های جایگزین در مناطق غنی از داده مانند مدیترانه قابل مقایسه بود. به‌عنوان مثال، (Vega-Garcia *et al.*, 2019) ۵ مورد از ۲۴۰ ایستگاه جریان‌سنج در حوزه آبریز Ebro که جریان طبیعی و بدون نقص را با دامنه داده‌های قابل اعتماد ۳۰ سال سوابق آب و هوا و رکوردهای جریان روزانه و با کمتر از سه گپ به دامنه $R = 0.7 - 0.8$ با یک مدل پیشرفته ANN رسیده‌اند. بنابراین، در این الگوریتم امکان شبیه‌سازی دقیق و قابل اعتماد داده‌های گمشده، به‌سرعت و به‌صورت خودکار فراهم و آن برای برنامه‌های کاربردی در مناطق کم‌داده مناسب تلقی می‌شود. هیدروگراف‌های بازسازی شده برای دوره ۲۰۱۸–۱۹۷۲ امکان تجزیه و تحلیل تغییر و تنوع در رژیم جریان‌های و تعامل آن با متغیرهای اقلیمی کلیدی مانند بارش و دما را در حوزه بلوچستان جنوبی فراهم می‌کند.

پیشنهاد می‌گردد، اثرات حوزه‌های آبریز متفاوت با ویژگی‌های هیدروفیزیکی– اقلیمی مشخص در مطالعات بعدی در عملکرد الگوریتم جنگل گمشد مورد تجزیه و تحلیل قرار گیرد. از جمله مسائل دیگر که لازم است در مطالعات آتی به آن‌ها پرداخته شود، بررسی روش پیشنهادی این مطالعه در مناطق اقلیمی و جغرافیایی دیگر، حساسیت‌سنجی به رژیم بارش و جریان و در نهایت بررسی عملکرد آن نسبت به سایر روش‌های متداول می‌باشد.

تشکر و قدردانی

از بازبینان ناشناس بابت ارائه نظرات سازنده در راستای بهبود کیفیت مقاله صمیمانه قدردانی می‌گردد. این پژوهش از طرف شرکت سهامی آب منطقه‌ای استان سیستان و بلوچستان در قالب پروژه تقاضا محور پشتیبانی شده که نویسندگان مراتب قدردانی خود را اعلام می‌دارند.

References

Aissia, M. A. B., Chebana, F., & Ouarda, T. B. (2017). Multivariate missing data in hydrology–Review and applications. *Advances in Water Resources*, *110*, 299-309.

Alibakhshi, S. M., Farid Hossini, A., Davari, K., Alizadeh, A., & Munyka, H. (2019). Assessment of Ground Station, GPM Satellite and MERRA Precipitation Products in Kashafrud Basin. *Watershed Management Research*, 9(18), 111-122 (In Persian).

Arriagada, P., Dieppois, B., Sidibe, M., & Link, O. (2019). Impacts of Climate Change and Climate Variability on Hydropower Potential in Data-Scarce Regions Subjected to Multi-Decadal Variability. *Energies*, 12, 2747.

[DOI: 10.61186/jwmr.15.2.49]

Bennett, D. A. (2001). How can I deal with missing data in my study? *Australian and New Zealand journal of public health*, *25*(5), 464-469.

Blum, A. G., Archfield, S. A., & Vogel, R. M. (2017). On the probability distribution of daily streamflow in the United States. *Hydrology and Earth System Sciences*, *21*(6), 3093-3103.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Damadi, S., Dehvari, A., Dahmardeh ghaleno, M. R., & Ebrahimiyan, M. (2021). Flood hazard zonation using HEC-RAS hydraulic model in Sarbaz River, Sistan and Baluchestan Province. *Watershed Engineering and Management*, *13*(3), 590-601 (In Persian).

Dembélé, M., Oriani, F., Tumbulto, J., Mariéthoz, G., & Schaefli, B. (2019). Gap-filling of daily streamflow time series using Direct Sampling in various hydroclimatic settings. *Journal of Hydrology*, *569*, 573-586.

Déry, S. J., Stahl, K., Moore, R. D., Whitfield, P. H., Menounos, B., & Burford, J. E. (2009). Detection of runoff timing changes in pluvial, nival, and glacial rivers of western Canada. *Water Resources Research*, *45*(4).

Deshmukh, H., Papageorgiou, M., Kilpatrick, E. S., Atkin, S. L., & Sathyapalan, T. (2019). Development of a novel risk prediction and risk stratification score for polycystic ovary syndrome. *Clinical Endocrinology*, *90*(1), 162-169.

Di Zio, M., Guarnera, U., & Luzi, O. (2007). Imputation through finite Gaussian mixture models. *Computational Statistics & Data Analysis*, *51*(11), 5305-5316.

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

Elshorbagy, A. A., Panu, U. S., & Simonovic, S. P. (2000). Group-based estimation of missing hydrological data: I. Approach and general methodology. *Hydrological sciences journal*, *45*(6), 849-866.

Grantham-McGregor, S., Cheung, Y. B., Cueto, S., Glewwe, P., Richter, L., & Strupp, B. (2007). Developmental potential in the first 5 years for children in developing countries. *The lancet*, *369*(9555), 60-70.

Gyau-Boakye, P., & Schultz, G. A. (1994). Filling gaps in runoff time series in West Africa. *Hydrological sciences journal*, *39*(6), 621-636.

Hamzah, F. B., Mohd Hamzah, F., Mohd Razali, S. F., Jaafar, O., & Abdul Jamil, N. (2020). Imputation methods for recovering streamflow observation: A methodological review. *Cogent Environmental Science*, *6*(1), 1745133.

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), 618-636.

Hawthorne, G., & Elliott, P. (2005). Imputing cross-sectional missing data: Comparison of common techniques. *Australian & New Zealand Journal of Psychiatry*, *39*(7), 583-590.

Heidari Chenari, F., Fazloula, R., & Nikzad Tehrani, E. (2022). Calibration and Evaluation of HEC-HMS Hydrological Model Parameters in Simulation of Single Rainfall-Runoff Events (Case Study: Tajan Watershed). *Watershed Management Research*, *13*(26), 69-81 (In Persian).

Hong, S., & Lynn, H. S. (2020). Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. *BMC medical research methodology*, *20*(1), 1-12.

Huisman, M. (2009). Imputation of missing network data: Some simple procedures. *Journal of Social Structure*, *10*(1), 1-29.

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., & Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. *Atmospheric environment*, *38*(18), 2895-2907.

Kanani, R., Fakheri Fard, A., Ghorbani, M. A., & Dinpashoh, Y. (2020). Trend Analysis of the Streamflow in the Lighvan River hydrometric Stations (Upstream and Downstream). *Watershed Management Research*, *11*(22), 11-19 (In Persian).

Kim, M., Baek, S., Ligaray, M., Pyo, J., Park, M., & Cho, K. H. (2015). Comparative studies of different imputation methods for recovering streamflow observation. *Water*, *7*(12), 6847-6860.

Kling, H., Fuchs, M., & Paulin, M. (2012). Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *Hydrology*, *424*, 264-277.

Knoben, W. J., Freer, J. E., & Woods, R. A. (2019). Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrology and Earth System Sciences*, *23*(10), 4323-4331.

Koçak, E. Prediction of daily fine particulate matter (PM2. 5) concentration in Aksaray, Turkey: Temporal variation, meteorological dependence, and employing artificial neural network. *Environmental Progress & Sustainable Energy*, e14355.

Lakshminarayan, K., Harp, S. A., & Samad, T. (1999). Imputation of missing data in industrial databases. *Applied intelligence*, *11*(3), 259-275.

Liu, J., & Zhang, Y. (2017). Multi-temporal clustering of continental floods and associated atmospheric circulations. *Journal of Hydrology*, *555*, 744-759.

Lopes, A. V., Chiang, J. C. H., Thompson, S. A., & Dracup, J. A. (2016). Trend and uncertainty in spatial-temporal patterns of hydrological droughts in the Amazon basin. *Geophysical Research Letters*, *43*(7), 3307-3316.

[DOI: 10.61186/jwmr.15.2.49]

Mackay, S. J., Arthington, A. H., & James, C. S. (2014). Classification and comparison of natural and altered flow regimes to support an Australian trial of the Ecological Limits of Hydrologic Alteration framework. *Ecohydrology*, *7*(6), 1485-1507.

Marino, S., Zhou, N., Zhao, Y., Wang, L., Wu, Q., & Dinov, I. D. (2019). HDDA: DataSifter: statistical obfuscation of electronic health records and other sensitive datasets. *Journal of statistical computation and simulation*, *89*(2), 249-271.

McGregor, G. R. (2019). Climate and rivers. *River Research and Applications*, *35*(8), 1119-1140.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., & Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, *50*(3), 885-900.

Muñoz, P., Orellana-Alvear, J., Willems, P., & Célleri, R. (2018). Flash-flood forecasting in an Andean Mountain catchment—development of a step-wise methodology based on the random forest algorithm. *Water*, *10*(11), 1519.

Nadi, M., Baziarpour, H., & Raeini sarjaz, M. (2022). Evaluation and modification of Aphrodite precipitation network in estimating monthly and annual precipitation in central parts of Iran. *Watershed Management Research*, *13*(25), 97-104 (In Persian).

Norazian, M. N., Shukri, Y. A., Azam, R. N., & Al Bakri, A. M. M. (2008). Estimation of missing values in air pollution data using single imputation techniques. *Science Asia*, *34*(3), 341-345.

Petrone, K. C., Hughes, J. D., Van Niel, T. G., & Silberstein, R. P. (2010). Streamflow decline in southwestern Australia, 1950–2008. *Geophysical Research Letters*, *37*(11).

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

Poff, N. L., Allan, J. D., Bain, M. B., Karr, J. R., Prestegaard, K. L., Richter, B. D., ... & Stromberg, J. C. (1997). The natural flow regime. *BioScience*, *47*(11), 769-784.

Sartori, N., Salvan, A., & Thomaseth, K. (2005). Multiple imputation of missing values in a cancer mortality analysis with estimated exposure dose. *Computational statistics & data analysis*, *49*(3), 937-953.

Schafer, J.L. (1997) *The Analysis of Incomplete Multivariate Data*. Chapman & Hall, London.

Sidibe, M., Dieppois, B., Mahé, G., Paturel, J. E., Amoussou, E., Anifowose, B., & Lawler, D. (2018). Trend and variability in a new, reconstructed streamflow dataset for West and Central Africa, and climatic interactions, 1950–2005. *Journal of hydrology*, *561*, 478-493.

Starrett, S.K., Heier, T., Su, Y., Bandurraga, M., Tuan, D., & Starrett, S. (2010). An example of the impact that filled-in peakflow data can have on flood frequency analysis, in: *Challenges of Change - Proceedings of the World Environmental and Water Resources Congress 2010*, 2451–2455.

Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112-118.

Tang, F., & Ishwaran, H. (2017). Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *10*(6), 363-377.

Tao, N., Chen, Y., Wu, Y., Wang, X., Li, L., & Zhu, A. (2019). The terpene limonene induced the green mold of citrus fruit through regulation of reactive oxygen species (ROS) homeostasis in Penicillium digitatum spores. *Food chemistry*, *277*, 414-422.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., ... & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, *17*(6), 520-525.

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), 910.

Ukkola, A. M., Keenan, T. F., Kelley, D. I., & Prentice, D. I. (2016). Vegetation plays an important role in mediating future water resources. *Environmental Research Letters*, *11*(9), 094022.

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical methods in medical research*, *16*(3), 219-242.

Vega-Garcia, C., Decuyper, M., & Alcázar, J. (2019). Applying cascade-correlation neural networks to in-fill gaps in Mediterranean daily flow data series. *Water*, 11(8), 1691.

Waljee, A. K., Mukherjee, A., Singal, A. G., Zhang, Y., Warren, J., Balis, U., Marrero, J., Zhu, J., & Higgins, P. D. (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ open*, 3(8), e002847.

Widaman, K. F. (2006). Best practices in quantitative methods for developmentalists: III. Missing data: What to do with or without them. *Monographs of the Society for Research in Child Development*, 7(1), 210-211.

Williams, L. S., Khosravi, B., Velimirovic, M., Khouri, J., Raza, S., Mazzoni, S., ... & Anwer, F. (2023). An Ensemble Machine Learning Model Using Gradient Boosting Identifies Patients with Disease Progression in Newly Diagnosed Multiple Myeloma. *Blood*, *142*, 1990.0.

Zhang, Y., & Post, D. (2018). How good are hydrological models for gap-filling streamflow data? *Hydrology and Earth System Sciences*, *22*(8), 4593-4604.

[DOI: 10.61186/jwmr.15.2.49]