International Scientific Conference 20 UNITECH 26

INTERNATIONAL SCIENTIFIC CONFERENCE
19-21 November 2026
GABROVO

EXPLAINABLE PREPROCESSING AND MISSING-VALUE BENCHMARKING OF DAILY HYDRO-METEOROLOGICAL TIME SERIES

Stefan Pitulić[1*], Dragana Radosavljević[1], Siniša Ilić[1], Vladimir Veljović[2]

[1]Faculty of Technical Sciences, Department of Computer Science and Informatics, University of Priština in Kosovska Mitrovica, Serbia
[2]Faculty of Technical Sciences Čačak, University of Kragujevac, Serbia
**Corresponding author: stefan.pitulic@pr.ac.rs*

Abstract
Daily hydro-meteorological records are often stored as wide Excel matrices in which rows represent years and consecutive blocks of 365 or 366 columns represent variables or stations. This paper presents an explainable workflow for detecting such blocks, classifying their physical type, distinguishing true missing values from calendar gaps, and preparing data for analysis. The web application uses a Python/FastAPI backend and a React frontend and supports linear interpolation, nearest-neighbour mean, day-of-year (DOY) median, and rank-preserving DOY best fit. In a controlled benchmark using 25 years of daily observations from 1991–2015, 4,113 values—5.00% of 82,179 eligible non-calendar cells—were removed as gaps of one to five days and reconstructed by all four methods. Block-wise full-range NRMSE identified linear interpolation as best in the three streamflow and three temperature blocks, whereas rank-DOY best fit performed best in both precipitation blocks and the relative-humidity block. Linear interpolation achieved the lowest or tied-lowest median block-wise NRMSE across all five gap lengths. A Friedman test across the nine blocks indicated an overall difference among methods ($p = 0.0133$). The calendar-aware workflow reduces manual preprocessing and supports transparent, variable-specific imputation-method selection.

Keywords: hydro-meteorological time series, missing data, imputation, explainable preprocessing, gap-length sensitivity

1. INTRODUCTION

Daily hydro-meteorological time series are fundamental for hydrology, meteorology, climate monitoring and environmental modelling. Numerous hydrological and hydro-meteorological models use such records as input data. Multisite streamflow simulation provides an example of modelling based on observations from several monitoring stations [1], while rainfall–runoff models commonly combine multiple meteorological inputs, such as precipitation and air temperature, to estimate discharge [2]. Consequently, before forecasting, statistical testing or model development, these records should be examined with respect to temporal variability, dependence among stations and variables, and the occurrence of missing observations. They must then be transformed into a clean and reproducible structure. This transformation constitutes data preprocessing, which is often one of the most demanding stages of practical analysis; its transparent automation can accelerate the workflow, facilitate result verification and support reliable repetition of the analysis.

In many engineering workflows daily data are stored as wide Excel matrices: rows represent years, while consecutive blocks of 365 or 366 columns represent variables, stations or derived quantities. Conceptually, each block is a contiguous group of daily columns, while isolated blank cells or short blank runs within a block represent candidate

This is an open access article licensed under Creative Commons Attribution 4.0 International
Print ISSN: 1313-230X, Online ISSN: 3033-1404

data gaps that require validation. Manual recognition of these blocks, their physical meaning and their missing-data structure is slow and error-prone.

The proposed software, ts-block-recognizer, automatically detects daily blocks and candidate data gaps, classifies each block as streamflow, precipitation, temperature or humidity, and allows the user to verify, rename and reorder the blocks interactively. The backend separates genuine missing observations from structural 29 February blanks and exports Metadata, Reordered_Matrix, Long_Format and Imputation_Log outputs. The deployed application is publicly available at https://tsrecognizer.pr.ac.rs/ [3]. To determine which imputation method most accurately reconstructs missing hydro-meteorological observations, a controlled benchmark was designed in which selected known values were randomly removed from the original matrix to create synthetic data gaps, while their original values and positions were stored separately in a truth table. Each method reconstructed the same incomplete matrix, after which the imputed values were compared with the preserved original values to calculate reconstruction errors. The present validation is limited to nine data blocks and one controlled 5% mixed-gap realization.

2. RELATED WORK

Missing values are a central issue in statistical data analysis. Classical literature distinguishes missing completely at random, missing at random and missing not at random mechanisms [4]. In environmental data, gaps may result from sensor failure, communication interruptions, maintenance periods or manual errors.

Recent reviews separate simple approaches, such as mean, median and interpolation, from advanced approaches based on ARIMA, K-nearest neighbours (KNN), multiple imputation, neural networks and generative models [5], [6]. Ahn et al. found that KNN can outperform simpler alternatives in some forecasting settings [7], whereas hydrological studies have reported benefits from regression and multiple-imputation strategies [8]. Multiple imputation by chained equations provides a principled framework for multivariate missing data [9]. K-nearest-neighbour imputation estimates a missing value from the observed values of the most similar records, which are selected according to a distance measure [10]. This approach differs from the simpler nearest-neighbour mean evaluated in the present study, which uses the closest valid observations in time within the same annual record. However, no method is universally optimal: performance depends on temporal smoothness, seasonality, intermittency of the variable, the frequency and magnitude of extreme values, site-specific characteristics, and the length and pattern of missing-data gaps.

Beyond classical statistical imputation, recent research has increasingly focused on data-driven reconstruction and standardized benchmarking. BRITS learns missing values directly through bidirectional recurrent temporal dynamics [11]. TSI-Bench standardizes missingness patterns, rates and paired evaluation settings across large collections of time-series imputation experiments [12]. ImputeGAP similarly emphasizes modular gap simulation, reproducible benchmarking, explainability and downstream evaluation [13]. These developments support the need for controlled and traceable evaluation protocols, while the present study focuses on transparent algorithms and calendar-aware hydro-meteorological matrices.

In contrast to frameworks centred primarily on advanced prediction models, the proposed workflow integrates structural recognition of wide 365/366-day matrices, explainable physical-variable classification, protected leap-day handling, interactive user validation and paired imputation benchmarking in a single publicly accessible application. The emphasis on human-verifiable evidence is consistent with the broader literature on interpretable model decisions [14]. The custom rank-preserving day-of-year (DOY) best-fit method is

positioned between local interpolation and distribution-based seasonal reconstruction, while the overall workflow automates preprocessing and missing-value imputation with interactive validation and traceable outputs.

3. INPUT RECOGNITION AND CALENDAR-AWARE GAPS

Let $X \in \mathbb{R}^{(N\times C)}$ denote the complete numerical matrix, where N is the number of annual records and C is the total number of daily data columns. The algorithm identifies a daily block width $W \in \{365,366\}$. When C is divisible by W, Equation (1) gives both the number of non-overlapping blocks and the column range assigned to each block.

The annual-record index r identifies a matrix row, y_r denotes the calendar year associated with row r, d is the DOY index, and $x_{r,d}$ denotes the observation recorded in annual record r on day d.

$$S = \frac{C}{W}, \quad B_s = X[:, (s-1)W + 1 : sW], \quad s = 1, \dots, S \tag{1}$$

Here, $X[r_1 : r_N, c_1 : c_M]$ denotes the submatrix of X comprising rows r_1 through r_N and columns c_1 through c_M. The symbol ":" without specified lower and upper bounds denotes all indices along the corresponding matrix dimension. Therefore, $X[:, (s-1)W + 1 : sW]$ selects all annual-record rows and the contiguous columns from $(s-1)W + 1$ to sW. The notation follows one-based indexing. Each block B_s therefore contains all daily observations of one physical variable or monitoring location across the available annual records.

In a 366-day block, day $d = 60$ represents 29 February. A blank value at this position in a non-leap year is an expected structural calendar gap rather than a data-quality error. Equation (2) formalizes this rule through a protected calendar mask, consistent with climatological practice for leap day [15]. Here, $1[\cdot]$ denotes the indicator function, which equals 1 when the condition in brackets is true and 0 otherwise.

$$M_{cal}(r,d) = 1, \left[\begin{array}{c} W = 366 \wedge d = 60 \\ \wedge \neg \mathrm{Leap}(y_r) \wedge \mathrm{isNaN}(x_{r,d}) \end{array}\right] \tag{2}$$

Once the protected positions have been identified, Eq. (3) isolates the blanks that represent genuine missing observations.

$$M_{real}(r,d) = 1, \quad [isNaN(xr,d)] \cdot [1 - M_{cal}(r,d)] \tag{3}$$

Consequently, only positions for which $M_{real}(r,d) = 1$ are reported as missing observations and offered for imputation. Protected leap-day blanks remain unchanged in every exported dataset. The overall system architecture and data flow are shown in Fig.1.

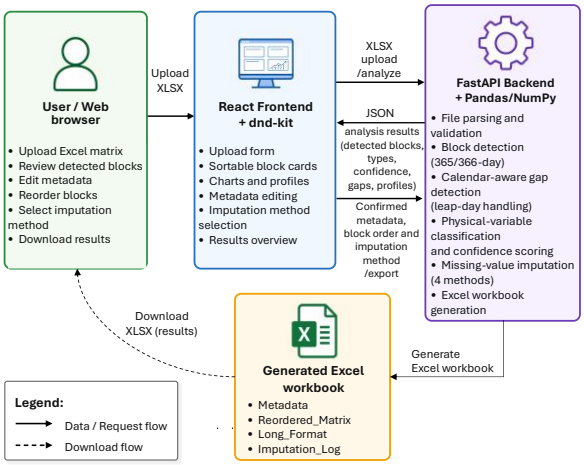


Fig. 1. *Architecture and data flow of the ts-block-recognizer system. The React frontend manages user interaction and block configuration, while the FastAPI backend performs matrix analysis, calendar-aware gap detection, classification, imputation and generation of the final Excel workbook.*

The backend is implemented with FastAPI [16] and exposes analysis and export services, while the React frontend [17] presents sortable cards, confidence explanations and seasonal profiles. User-confirmed metadata and imputation logs are preserved in the exported files.

4. EXPLAINABLE CLASSIFICATION AND CONFIDENCE SCORE

Each block is represented by interpretable features, including the positive-value ratio, zero ratio, negative-value ratio, skewness, lag-1 autocorrelation, coefficient of variation, seasonal amplitude and the proportion of values within the interval 0–100. The ramp transformation in Eq. (4) places every feature on [0,1], so its contribution increases gradually rather than through a hard threshold.

$$\text{ramp}(x;a,b) = \begin{cases} 0, & x \le a \\ \dfrac{x-a}{b-a}, & a < x < b \\ 1, & x \ge b \end{cases} \quad (4)$$

Parameters a and b determine the interval over which the contribution of a feature gradually changes from no evidence to full evidence.

For each candidate physical-variable class c in the set {streamflow, precipitation, temperature, relative humidity}, the normalized features are combined into the weighted evidence score Ec(Bs) for block Bs, as defined in Eq. (5).

$$E_c(B_s) = \sum_{i=1}^{p} w_{i,c}\, f_i(B_s) \quad (5)$$

Here, $f_i(B_s)$ is the normalized value of feature i, $w_{(i,c)}$ is its predefined weight for class c, and p is the number of extracted features. The classes are not features; each class is supported by a characteristic feature pattern. Streamflow is supported by positive and continuous values with few zeros; precipitation by many zeros and positive skewness; temperature by seasonal amplitude, negative values and temporal smoothness; and relative humidity by the 0–100 range and a high proportion of values above 50%.

Equation (6) converts the class-specific evidence scores into normalized probabilities by means of a temperature-scaled softmax.

$$P(c \mid B_s) = \frac{\exp[\tau E_c(B_s)]}{\sum_{k=1}^{K} \exp[\tau E_k(B_s)]} \quad (6)$$

The implementation uses the empirically selected softmax-sharpening value $\tau = 6.5$. This value is not derived from a physical law or formal optimization; it sharpens the normalized probability distribution and increases the separation between a clearly supported class and weaker alternatives. Confidence and the separation margin are then obtained from Eqs. (7a) and (7b), respectively.

$$Confidence(B_s) = max_c P(c|B_s) \quad (7a)$$
$$Margin(B_s) = P(c_1|B_s) - P(c_2|B \quad (7b)$$

Classes c_1 and c_2 denote the highest- and second-highest-ranked candidates. The interface additionally presents the three strongest candidates and the feature-based explanation of the leading classification.

5. IMPUTATION ALGORITHMS

The software does not fill data automatically. Once true gaps are detected, the user selects one of four methods. Calendar gaps remain protected and are excluded from all algorithms.

5.1. Linear interpolation

Linear interpolation assumes that the values inside a missing run change uniformly between the nearest valid observations that bound the gap. Equation (8) places the L missing points at equally spaced positions between the two bounding observations x_a and x_b.

$$\hat{x}_{a+k} = x_a + \frac{k}{L+1}(x_b - x_a), \quad k = 1, \dots, L \quad (8)$$

This method is expected to perform well for short gaps in smooth series, but it may attenuate abrupt peaks in intermittent variables such as precipitation.

5.2. Nearest-neighbour mean

The nearest-neighbour mean method searches within the same annual record for the closest valid observations to the left and right of a missing position. Its reconstruction

and one-sided fallback cases are collected in Eq. (9).

$$\hat{x}_t = \begin{cases} \frac{x_l + x_r}{2}, & x_l \text{ and } x_r \text{ are availab} \\ x_l, & x_r \text{ is unavailable} \\ x_r, & x_l \text{ is unavailable} \end{cases} \quad (9)$$

For a multi-day gap, all positions may therefore depend strongly on the same two boundary observations, which can reduce accuracy as the gap becomes longer.

5.3. DOY median

The DOY median method uses observations from the same day of year in the remaining annual records. For year y and day d, Eq. (10) forms the estimate from valid observations at the same day of year in all other records.

$$\hat{x}_{y,d} = \text{median}\{x_{y',d} : y' \neq y, x_{y',d} \text{ is valid}\} \quad (10)$$

If no same-day donor is available, the implementation first uses the median of the affected annual record and then the global median of the block. The method is robust to extreme donor values, but it does not explicitly account for the local conditions of the affected year.

5.4. Rank-preserving DOY best fit

Rank-preserving DOY best fit combines the local temporal context of the affected year with the empirical distribution of the same day of year across the remaining records. The blend in Eq. (11) combines the donor quantile with the local expectation.

$$\hat{x}_{y,d} = \lambda Q_d(q_y) + (1-\lambda)e_{y,d}, \quad \lambda = 0.65 \quad (11)$$

Here, $e_{y,d}$ is the local expectation derived from neighbouring valid days for year y and day d; $Q_d(q_y)$ is the empirical donor quantile for day d; and q_y is assigned from the rank of $e_{y,d}$ among missing cases for the same day. This formulation preserves both the relative ordering of reconstructed cases and the day-of-year distribution. The empirical value $\lambda = 0.65$ assigns slightly greater weight to the day-of-year donor quantile than to the local expectation; it is not obtained through formal optimization.

6. CONTROLLED MISSING-DATA BENCHMARK

The case study comprises 25 years of daily hydro-meteorological observations covering the period 1991–2015. The streamflow series were obtained from the gauging stations Devići, Mlanča and Ušće and were assigned to blocks B001, B002 and B003, respectively. Daily precipitation records from the meteorological stations Beograd and Zlatibor were assigned to blocks B004 and B005. The remaining four blocks contain meteorological observations from the Beograd station: maximum air temperature T_{max} in B006, minimum air temperature T_{min} in B007, mean air temperature T_{mean} in B008, and mean relative humidity RH_{mean} in B009. Together, these series provide 25 annual rows for each of nine 366-column data blocks. Streamflow is expressed in $m^3 s^{-1}$, precipitation in mm, air temperature in $°C$, and relative humidity in %.

The streamflow records used for blocks B001–B003 were obtained from the publicly available Hydrological Yearbooks – Surface Waters published by the Republic Hydrometeorological Service of Serbia (RHMSS) [18]. The precipitation, temperature and relative-humidity records used for blocks B004–B009 were obtained from the RHMSS Meteorological Yearbooks – Climatological Data [19]. Both collections provide direct, reproducible access to the annual source records.

The controlled benchmark followed a paired ground-truth design, with the matrix construction, random gap generation, preservation of removed values and exact reconstruction matching described in Sections 6.1 and 6.2.

The benchmark started from a complete matrix containing 25 annual records and 3,294 daily columns, corresponding to 82,350 cells organized into nine 366-day blocks. The original matrix contained 171

structural leap-day blanks, leaving 82,179 valid positions eligible for controlled deletion. A 5% deletion rate, corresponding to 4,113 positions, was selected as a practical compromise because it provides enough missing values for reliable method comparison while limiting disturbance of the original dataset's main statistical characteristics. The complete matrix was copied before any values were removed. Synthetic non-overlapping gaps of one to five consecutive days were then introduced independently within each data block. Before a selected value was replaced by NaN, its original value and positional identifiers were stored in an authoritative truth table. The resulting incomplete matrix was reconstructed independently by all four algorithms.

6.1. Controlled deletion and ground-truth construction

Let X_{full} denote the original complete matrix. The admissible set Ω in Eq. (12) contains only finite observations outside the protected calendar mask and therefore excludes both original gaps and structural leap-day blanks.

$$\Omega = \{(b,r,d): X_{full}(b,r,d) \in \mathbb{R} \wedge M_{cal}(b,r,d) = 0\} \quad (12)$$

Here, b indexes the data block, r indexes the annual record (calendar year), and d indexes the day of year. Xfull(b,r,d) denotes the original value, and Mcal(b,r,d) is the protected calendar mask.

A separate matrix $X_{miss} = X_{full}$ is created, and only X_{miss} is modified. Independent, non-overlapping gaps with lengths $L \in \{1,2,3,4,5\}$ are sampled randomly within each block. Every removed observation is entered in truth table T before it is replaced by NaN. Equation (13) then calculates the realized deletion rate directly from that truth table rather than inferring it from the configuration parameter.

$$p_{\mathrm{del}} = 100\frac{|T|}{|\Omega|} = 100\frac{4113}{82179} = 5.00\% \quad (13)$$

The truth table contained 4,113 unique positions: 252 values from one-day gaps, 562 from two-day gaps, 915 from three-day gaps, 1,064 from four-day gaps and 1,320 from five-day gaps. All 4,113 positions were reconstructed by every method and entered the metric calculations. The truth table stores the scenario, replicate, block identifier, annual-record index, year, day of year, source column, gap length and exact removed value.

6.2. Paired imputation and exact value matching

The same incomplete matrix X_{miss} and the same deletion mask were used for all four methods. It was uploaded once to ts-block-recognizer and exported four times, using linear interpolation, nearest-neighbour mean, DOY median and rank-DOY best fit. The truth table and hidden original values were not supplied to the imputation software, thereby preventing information leakage. The independent evaluation routine subsequently matched each reconstructed value to its stored original value using the block identifier, annual-record index and day of year.

For imputation method m and deleted position i, Eq. (14) expresses the signed point-wise reconstruction error as the difference between the imputed value and the stored original value.

$$e_{i,m} = \hat{y}_{i,m} - y_i, \qquad i \in T \quad (14)$$

All evaluation metrics are calculated exclusively from positions contained in T. Originally missing cells, protected calendar blanks and observations that were not synthetically deleted do not contribute to the benchmark scores.

6.3. Aggregation, pairing and reproducibility safeguards

The errors from Eq. (14) are aggregated primarily by individual data block and by consecutive-gap length within each block. Raw MAE and RMSE are calculated and interpreted only within blocks, where all values share the same physical unit, whereas cross-block comparisons use the full-range

NRMSE defined in Eq. (15c). Since every method is evaluated on exactly the same truth-table positions, all method comparisons are paired. Gap locations are independent across blocks, representing local station or sensor outages rather than a synchronized system-wide interruption.

The random seed, benchmark configuration, detected block metadata, protected calendar mask and complete truth table are preserved together with the evaluation outputs. Numerical summaries, point-wise errors, statistical-test results and figure source data are exported, enabling the experiment to be audited and independently repeated. The realized deletion rate is calculated directly from the truth table and checked against the nominal configuration value. The complete controlled benchmark procedure is summarized in Algorithm 1.

Algorithm 1. Reproducible benchmark workflow

1. Read X_{full} and detect the block width W using Eq. (1).
2. Construct Ω using Eqs. (2), (3) and (12).
3. Copy X_{full} to X_{miss} and sample independent, non-overlapping gaps of one to five days.
4. Store every removed value and its identifiers in truth table T.
5. Reconstruct X_{miss} independently using each of the four methods.
6. Match reconstructed and original values and calculate Eq. (14).
7. Aggregate block-wise metrics, normalize cross-block errors, generate paired tests, numerical tables.

7. EVALUATION METRICS

For each data block b, let $T_b \subset T$ denote the subset of truth-table positions belonging to that block, and let $M_b = |T_b|$ denote the number of deliberately deleted and successfully reconstructed positions in that block. For $i \in T_b$, the stored original values y_i are compared with the reconstructed values $\hat{y}_i$. The suite of metrics in Eqs. (15a)–(15i) measures complementary aspects of reconstruction quality. Lower MAE, RMSE, NRMSE and absolute bias indicate better performance, whereas higher R^2, NSE, KGE and Willmott d indicate better agreement.

$$MAE = \frac{1}{M_b}\sum_{i\in T_b}|y_i - \hat{y}_i| \tag{15a}$$

$$RMSE = \sqrt{\frac{1}{M_b}\sum_{i\in T_b}(y_i - \hat{y}_i)^2} \tag{15b}$$

$$NRMSE_{b,m}^{full} = \frac{\mathrm{RMSE}_{b,m}}{\max\left(X_b^{\mathrm{full}}\right) - \min\left(X_b^{\mathrm{full}}\right)} \tag{15c}$$

$$Bias = \frac{1}{M_b}\sum_{i\in T_b}(\hat{y}_i - y_i) \tag{15d}$$

$$PBIAS = 100\frac{\sum_{i\in T_b}(\hat{y}_i - y_i)}{\sum_{i\in T_b} y_i} \tag{15e}$$

$$R^2 = \mathrm{corr}(y, \hat{y})^2 \tag{15f}$$

$$NSE = 1 - \frac{\sum_{i\in T_b}(y_i - \hat{y}_i)^2}{\sum_{i\in T_b}(y_i - \bar{y})^2} \tag{15g}$$

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \tag{15h}$$

$$d = 1 - \frac{\sum_{i\in T_b}(y_i - \hat{y}_i)^2}{\sum_{i\in T_b}(|\hat{y}_i - \bar{y}| + |y_i - \bar{y}|)^2} \tag{15i}$$

The error and bias metrics are given in Eqs. (15a)–(15e), whereas the association and efficiency measures are provided in Eqs. (15f)–(15i). Under the sign convention used in Eq. (15e), negative PBIAS indicates underestimation, whereas positive PBIAS indicates overestimation.

Within the KGE expression in Eq. (15h), r is the Pearson correlation coefficient, $\alpha = \sigma_{\hat{y}}/\sigma_y$ is the variability ratio and $\beta = \mu_{\hat{y}}/\mu_y$ is the bias ratio, where $\hat{y}$ and y denote the reconstructed and original values, respectively. MAPE was not included in the final ranking because zero and near-zero streamflow and precipitation values can produce unstable or excessively large percentage errors. NSE, KGE and Willmott d are widely used in hydrological evaluation [20]–[23].

To enable comparison across variables with different physical units and numerical scales, NRMSE in Eq. (15c) was normalized

by the range of all valid observations in the corresponding original complete block. Here, X_b^{full} denotes all valid observations in original complete block b. This denominator is fixed across the four methods and does not depend on the particular observations selected for controlled deletion. NRMSE was used for cross-block comparison in Fig. 2, for the gap-length analysis in Fig. 3, and for the non-parametric statistical comparison. MAE and RMSE from Eqs. (15a) and (15b) were interpreted only within individual blocks in their native physical units.

The non-parametric statistical comparison was based on the nine block-wise full-range NRMSE values obtained for each method. A Friedman test was used to assess whether the four methods differed across the nine paired data blocks. Pairwise Wilcoxon signed-rank tests were then applied as a post-hoc analysis, with Holm correction for the six pairwise comparisons.

8. RESULTS AND ANALYSIS

The benchmark comprised 25 annual records covering 1991–2015, arranged into nine 366-day blocks. The recognition module assigned the expected physical type to all blocks, with confidence scores ranging from 90.7% to 99.0%, while all 4,113 deliberately deleted values were successfully reconstructed by each method.

Table 1 reports block-wise MAE and RMSE in the native physical unit of each data block, while Table 2 summarizes the dimensionless full-range NRMSE rankings across blocks. Linear interpolation achieved the lowest full-range NRMSE in six of the nine blocks (B001–B003 and B006–B008), whereas rank-DOY best fit was best in B004, B005 and B009. Across the nine blocks, median full-range NRMSE values were 0.0590 for linear interpolation, 0.0642 for nearest-neighbour mean, 0.0651 for rank-DOY best fit and 0.0778 for DOY median.

Table 1. *Block-wise MAE and RMSE for the four imputation methods. Each cell reports MAE (top) and RMSE (bottom) in the native physical unit of the corresponding block.*

Block / unit	Method	MAE	RMSE
B001 Q m³/s	Linear	0.344	0.713
B001 Q m³/s	NN mean	0.377	0.745
B001 Q m³/s	Rank-DOY	0.829	1.385
B001 Q m³/s	DOY median	1.195	1.913
B002 Q m³/s	Linear	1.030	2.841
B002 Q m³/s	NN mean	1.412	3.729
B002 Q m³/s	Rank-DOY	1.775	4.105
B002 Q m³/s	DOY median	2.361	4.910
B003 Q m³/s	Linear	1.044	2.327
B003 Q m³/s	NN mean	1.226	2.579
B003 Q m³/s	Rank-DOY	2.169	3.779
B003 Q m³/s	DOY median	3.013	5.050
B004 P mm	Linear	2.442	5.395
B004 P mm	NN mean	2.390	5.286
B004 P mm	Rank-DOY	1.883	4.901
B004 P mm	DOY median	1.682	4.945
B005 P mm	Linear	3.444	6.150
B005 P mm	NN mean	3.283	5.978
B005 P mm	Rank-DOY	2.703	5.751
B005 P mm	DOY median	2.524	6.113
B006 Tmax °C	Linear	2.478	3.262
B006 Tmax °C	NN mean	2.957	3.843
B006 Tmax °C	Rank-DOY	3.559	4.360
B006 Tmax °C	DOY median	4.359	5.342
B007 Tmin °C	Linear	1.983	2.512
B007 Tmin °C	NN mean	2.082	2.734
B007 Tmin °C	Rank-DOY	2.616	3.234
B007 Tmin °C	DOY median	3.148	3.865
B008 Tmean °C	Linear	2.124	2.800
B008 Tmean °C	NN mean	2.283	3.082
B008 Tmean °C	Rank-DOY	3.067	3.701
B008 Tmean °C	DOY median	3.816	4.597
B009 RH %	Linear	8.999	11.886
B009 RH %	NN mean	9.907	12.982
B009 RH %	Rank-DOY	9.279	11.745
B009 RH %	DOY median	10.567	13.190

Table 2. *Cross-block summary of full-range NRMSE. Lower median NRMSE and mean rank indicate better performance; "best blocks" counts the blocks in which a method achieved the lowest NRMSE.*

Method	Median NRMSE	Mean rank	Wins
Linear	0.0590	1.78	6/9
NN mean	0.0642	2.22	0/9
Rank-DOY	0.0651	2.33	3/9
DOY median	0.0778	3.67	0/9

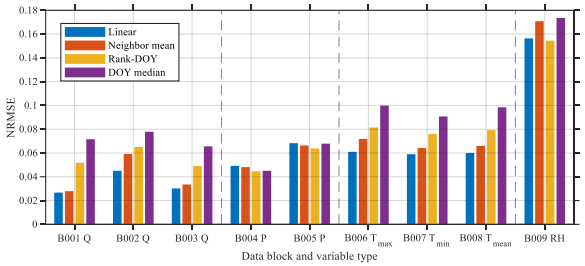


***Fig. 2**. Block-wise full-range NRMSE normalized by the range of valid observations in each original block. B001–B003: streamflow; B004–B005: precipitation; B006–B008: maximum, minimum and mean air temperature; B009: mean relative humidity. Lower values indicate better reconstruction accuracy.*

The normalized gap-length analysis in Fig. 3 shows that linear interpolation tied with nearest-neighbour mean for the lowest median block-wise NRMSE for one-day gaps (0.0420 for both) and achieved the lowest median NRMSE for gaps of two to five days. For five-day gaps, the median NRMSE values were 0.0592 for linear interpolation, 0.0702 for nearest-neighbour mean, 0.0776 for rank-DOY best fit and 0.0874 for DOY median. Thus, after errors are normalized within each block, no five-day crossover in favour of rank-DOY is observed.

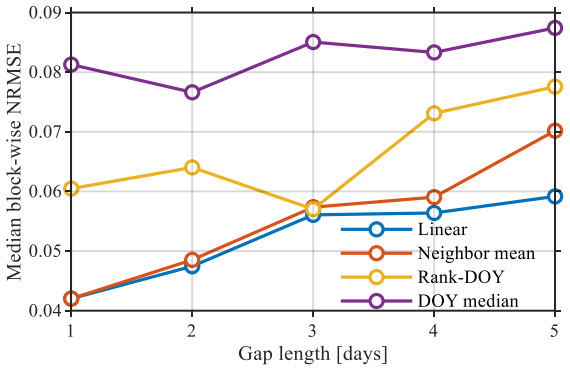


***Fig. 3**. Dependence of median block-wise full-range NRMSE on consecutive-gap length. NRMSE was calculated separately for each block using the range of valid observations in the corresponding original block and then summarized by the median across the nine blocks.*

The Friedman test indicated an overall difference among the four methods across the nine block-wise NRMSE values ($\chi^2(3) = 10.733$, $p = 0.0133$). After Holm correction, only the comparison between rank-DOY best fit and DOY median remained significant (adjusted $p = 0.0234$); all other adjusted pairwise p-values were at least 0.0977.

8.1. Practical method-selection summary

The numerical results support the following operational decision rules, which depend on both the physical variable and the expected gap duration:

- Streamflow and temperature: Linear interpolation is the preferred method, achieving the lowest block-wise NRMSE for six of the nine analysed blocks and benefiting from strong day-to-day continuity.
- Precipitation blocks B004–B005: Rank-DOY best fit achieved the lowest block-wise NRMSE for both precipitation blocks.
- Relative humidity and gap-length behaviour: Rank-DOY best fit was slightly more accurate for B009, while linear interpolation tied with nearest-neighbour mean for one-day gaps and achieved the lowest median block-wise NRMSE for gaps of two to five days.
- Simple transparent baseline: Nearest-neighbour mean is competitive for short gaps but degrades when the same boundary values represent a long gap.
- Seasonal robust fallback: DOY median is robust to outliers but is weakest overall because it ignores conditions specific to the affected year.

9. DISCUSSION

The block-wise and gap-length patterns are consistent with the assumptions underlying the four algorithms. Local interpolation benefits smooth and temporally continuous streamflow and temperature series. Rank-DOY best fit was more effective for the two precipitation blocks and relative humidity, showing the value of day-of-year donor information for these variable types. However, after normalizing errors within each block, linear interpolation tied with nearest-neighbour mean at one-day gaps and retained the lowest median NRMSE for gaps of two to five days. The weaker overall performance of the DOY median indicates that calendar position alone may be insufficient when conditions differ substantially among years.

These findings are consistent with recent benchmarking studies [12], [13], which emphasize that method rankings depend strongly on the missingness pattern, data domain and evaluation protocol. Unlike learning-based models such as BRITS [11], the four methods evaluated here require no model training and remain directly auditable.

The benchmark therefore does not support a universal winner across variable types. Linear interpolation is the strongest general baseline and performed best in six of nine blocks, whereas rank-DOY best fit is preferable for the two precipitation blocks and slightly better for relative humidity.

9.1. Limitations and threats to validity

The experiment used 25 annual records, one controlled 5% mixed-gap realization containing 4,113 deliberately deleted and successfully reconstructed positions, and independent synthetic gaps within each data block. Real outages may be temporally clustered, station-wide, synchronized across variables or related to extreme weather; therefore, the present results primarily represent a controlled random-deletion benchmark with local block-specific consecutive gaps. Because the analysis is based on a single mixed-gap realization, the stability of the method rankings across different random deletion patterns was not quantified. Repeated experiments using multiple random seeds, additional stations and years, and different missingness rates and patterns would enable the calculation of confidence intervals and provide a more reliable assessment of external validity.

MAE and RMSE were calculated and interpreted separately within each data block so that their native physical units were not mixed. NSE, KGE and Willmott d were likewise evaluated within blocks, while cross-block comparisons and statistical testing used the full-range NRMSE from Eq. (15c). The paired design included the same 457 scored positions per block and method. This strategy avoids combining raw errors with incompatible physical units, although range-based normalization may still be influenced by extreme values. Because the classification component was evaluated on only nine expected block labels, the reported confidence scores should be interpreted as internal relative support rather than externally calibrated probabilities.

10. CONCLUSION

This paper presents an explainable workflow for recognizing, validating and imputing wide daily hydro-meteorological matrices. The software correctly handled the 365/366-day structure, protected structural leap-day blanks, assigned the expected physical type to all nine analysed data blocks with high confidence, and provided traceable exports. In the controlled 5.00% mixed-gap benchmark comprising 4,113 scored positions from 25 annual records, block-wise full-range NRMSE identified linear interpolation as best in six blocks (all streamflow and temperature blocks), while rank-DOY best fit performed best in both precipitation blocks and relative humidity. Linear interpolation also achieved the lowest or tied-lowest median block-wise NRMSE across all five gap lengths. The combined web-based preprocessing and independent benchmark-evaluation workflow substantially reduces manual data preparation and supports data-driven selection of an imputation method before further modelling.

The traceable deletion procedure and preserved evaluation outputs improve transparency and facilitate independent verification and extension of the workflow. Future work will assess ranking stability across multiple random seeds, stations, years, and missingness scenarios and will extend the comparison to KNN, ARIMA, multiple-imputation, and learning-based baselines.

DATA AND SOFTWARE AVAILABILITY

The daily records analysed in this study cover the period 1991–2015 and are publicly available through the RHMSS Hydrological Yearbooks – Surface Waters [18] and the RHMSS Meteorological Yearbooks – Climatological Data [19]. The deployed ts-

block-recognizer application used for preprocessing and imputation is publicly accessible at https://tsrecognizer.pr.ac.rs/ [3]. The benchmark configuration, controlled-deletion records, and evaluation scripts are available from the corresponding author upon reasonable request.

ACKNOWLEDGMENTS:The authors would like to thank the Ministry of Science, Technological Development and Innovation of the Republic of Serbia for funding the scientific research work, contract no. 451-03-34/2026-03/200155, realized by the Faculty of Technical Sciences in Kosovska Mitrovica, University of Priština. The funder had no role in the study design, data preparation, analysis, interpretation of the results, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

S.P.: Conceptualization, methodology, software, data curation, formal analysis, investigation, validation, visualization, benchmark design, interpretation of results, writing—original draft, and writing—review and editing.
D.R.: Methodology, data curation, validation, interpretation of results, and writing—review and editing.
S.I.: Methodology, supervision, project coordination, validation, interpretation of results, and writing—review and editing.
V.V.: Validation and software testing. All authors reviewed and approved the final manuscript.

REFERENCES

[1] Pitulić S., Radosavljević D., Marković Đ., Ilić S. A hybrid rank-preserving and evolutionary algorithm for multisite daily streamflow simulation. Algorithms 2026;19(7):541. doi:10.3390/a19070541.

[2] Kratzert F., Klotz D., Brenner C., Schulz K., Herrnegger M. Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. Hydrology and Earth System Sciences 2018;22:6005–6022. doi:10.5194/hess-22-6005-2018.

[3] TS Block Recognizer web application. Available at: https://tsrecognizer.pr.ac.rs/ (accessed 10.07.2026).

[4] Little R.J.A., Rubin D.B. Statistical analysis with missing data. 3rd ed. Hoboken, NJ: John Wiley & Sons, 2019. doi:10.1002/9781119482260.

[5] Jaradat Y., Masoud M., Manasrah A., Alia M., Jannoud I. Review of data imputation techniques in time series data: comparative analysis. The Eurasia Proceedings of Science, Technology, Engineering & Mathematics 2024;27:122–129. doi:10.55549/epstem.1518433.

[6] Alejo-Sanchez L.E. et al. Missing data imputation of climate time series: a review. MethodsX 2025;15:103455. doi:10.1016/j.mex.2025.103455.

[7] Ahn H., Sun K., Kim K.P. Comparison of missing data imputation methods in time series forecasting. Computers, Materials & Continua 2022;70(1):767–779. doi:10.32604/cmc.2022.019369.

[8] Hamzah F.B., Mohd Hamzah F., Mohd Razali S.F., Samad H. A comparison of multiple imputation methods for recovering missing data in hydrological studies. Civil Engineering Journal 2021;7(9):1608–1619. doi:10.28991/cej-2021-03091747.

[9] van Buuren S., Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. Journal of Statistical Software 2011;45(3):1–67. doi:10.18637/jss.v045.i03.

[10] Troyanskaya O. et al. Missing value estimation methods for DNA microarrays. Bioinformatics 2001;17(6):520–525. doi:10.1093/bioinformatics/17.6.520.

[11] Cao W., Wang D., Li J., Zhou H., Li L., Li Y. BRITS: bidirectional recurrent imputation for time series. Advances in Neural Information Processing Systems 31, 2018, p. 6775–6785. Available at: https://proceedings.neurips.cc/paper/2018/hash/734e6bfcd358e25ac1db0a4241b95651-Abstract.html (accessed 11.07.2026).

[12] Du W. et al. TSI-Bench: benchmarking time series imputation. arXiv preprint arXiv:2406.12747, 2024. doi:10.48550/arXiv.2406.12747. Available at: https://arxiv.org/abs/2406.12747 (accessed 11.07.2026).

[13] Nater Q., Khayati M., Pasquier J. ImputeGAP: a comprehensive library for time series imputation. arXiv preprint arXiv:2503.15250, 2025. doi:10.48550/arXiv.2503.15250. Available at: https://arxiv.org/abs/2503.15250 (accessed 11.07.2026).

[14] Ribeiro M.T., Singh S., Guestrin C. Why should I trust you? Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, p. 1135–1144. doi:10.1145/2939672.2939778.

[15] National Centers for Environmental Information (NOAA NCEI). Hang on, what's not normal about February 29? 2024. Available at: https://www.ncei.noaa.gov/news/hang-on-whats-not-normal-about-February-29 (accessed 11.07.2026).

[16] FastAPI. FastAPI documentation. Available at: https://fastapi.tiangolo.com/ (accessed 08.07.2026).

[17] React. React documentation. Available at: https://react.dev/ (accessed 08.07.2026).

[18] Republic Hydrometeorological Service of Serbia (RHMSS). Hydrological yearbooks – surface waters. Available at: https://www.hidmet.gov.rs/ciril/hidrologija/povrsinske_godisnjaci.php (accessed 10.07.2026).

[19] Republic Hydrometeorological Service of Serbia (RHMSS). Meteorological yearbook – climatological data. Available at: https://www.hidmet.gov.rs/latin/meteorologija/klimatologija_godisnjaci.php (accessed 10.07.2026).

[20] Nash J.E., Sutcliffe J.V. River flow forecasting through conceptual models part I – a discussion of principles. Journal of Hydrology 1970;10(3):282–290. doi:10.1016/0022-1694(70)90255-6.

[21] Gupta H.V., Kling H., Yilmaz K.K., Martinez G.F. Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. Journal of Hydrology 2009;377(1–2):80–91. doi:10.1016/j.jhydrol.2009.08.003.

[22] Willmott C.J. On the validation of models. Physical Geography 1981;2(2):184–194. doi:10.1080/02723646.1981.10642213.

[23] Moriasi D.N., Arnold J.G., Van Liew M.W., Bingner R.L., Harmel R.D., Veith T.L. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Transactions of the ASABE 2007;50(3):885–900. doi:10.13031/2013.23153.