Journal of Big Data

RESEARCH Open Access

Disaggregating IMERG satellite precipitation over Czech Republic: an innovative approach using hybrid Extreme Gradient Boosting based on Fuzzy Spatial-Temporal Multivariate Clustering

Ujjwal Singh[1*], Sadaf Nasreen[1], Gaurav Tripathi[2], Pragya Mehrishi[3], Rajani Kumar Pradhan[1], Poppová Bestakova[1], Vivek Vikram Singh[4], K C Gouda[5], Laxmi Kant Sharma[6], Kiran Jalem[7], Petr Maca[1], Rama Rao Nidamanuri[8], Akhilesh Singh Raghubanshi[9], Yannis Markonis[1], Rakovec Oldřich[1] and Martin Hanel[1]

*Correspondence:
singh@fzp.czu.cz

Full list of author information is available at the end of the article

Abstract

Accurate precipitation estimation at high spatial and temporal resolutions is essential for hydrological and meteorological applications, especially in regions experiencing water resource degradation. This study presents a robust non-parametric framework for disaggregating coarse-resolution satellite precipitation data to finer scales, using a hybrid model that integrates Extreme Gradient Boosting (XGBoost) with multivariate spatio-temporal fuzzy clustering. Eight clusters were delineated based on Integrated Multi-satellite Retrievals for GPM (IMERG) precipitation and Shuttle Radar Topography Mission (SRTM) elevation data, with one representative station per cluster used for training and validation, and an additional 19 stations employed solely for independent validation. We downscaled 255 months (June 2000–September 2021) of IMERG precipitation data from 11 to 1 km spatial resolution across the Czech Republic. The disaggregated precipitation demonstrated marked accuracy improvements when evaluated against observed station data, with R^2 values ranging from 0.63 to 0.85, RMSE between 17.43 mm and 32.41 mm, NSE from 0.39 to 0.82, and KGE spanning 0.67 to 0.86-indicating a significant reduction in the bias inherent in the original IMERG data. The proposed methodology achieved (1) enhanced agreement between disaggregated and observed monthly precipitation, (2) significant improvement in IMERG data accuracy at finer scales, and (3) demonstrated operational potential in regions with sparse ground-based observations. This approach offers a promising solution for generating reliable, high-resolution precipitation datasets in data-scarce environments, with broad applicability in global hydrological and meteorological modelling.

Keywords: Disaggregation, Multivariate spatial-temporal fuzzy clustering, Extreme Gradient Boosting, Czech Republic

Introduction

Understanding precipitation variability is fundamental for evaluating hydro-meteorological extremes [1], disaster risk [2], irrigation management, water resource allocation [3], agricultural productivity [4], and hydrological modeling [5, 6]. However, accurate precipitation estimation remains challenging due to the sparse distribution and limited density of ground-based rain gauge networks, particularly in remote and topographically complex areas [7]. Rain gauges provide point-scale observations, often failing to represent broader regional spatial variability [8]. Consequently, satellite-based precipitation products have emerged as valuable alternatives.

Several satellite-derived precipitation products, including TRMM, CMORPH, CHIRPS, and PERSIANN, have significantly advanced hydrological and climatological research. Among these, the Integrated Multi-satellite Retrievals for GPM (IMERG) is particularly notable for its comprehensive global coverage and superior temporal resolution [9–11]. Nevertheless, the spatial resolution of IMERG data ($0.1° \times 0.1°$, approximately 11 km at the equator) is insufficient for capturing fine-scale variability, particularly in regions characterised by complex topography and heterogeneous land-surface characteristics [12]. Enhancing IMERG's spatial resolution through effective disaggregation methods is crucial for various local-scale hydrological applications, such as water resource planning, drought risk assessment, precision agriculture, and catchment-level hydrological modelling.

Numerous precipitation disaggregation methods have been explored to overcome spatial resolution limitations by employing regression techniques with appropriate predictors. Regression-based approaches, including exponential regression (ER), multiple linear regression (MLR), and geographically weighted regression (GWR), commonly use predictors like Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), Land Surface Temperature (LST), and elevation [13–15]. Despite their simplicity, ER and MLR often fail to accurately capture non-linear relationships and spatial heterogeneity, primarily due to their stationary assumptions [16]. GWR partially addresses these issues by allowing local model calibration but remains sensitive to multicollinearity among predictors [15].

Vegetation indices such as NDVI and EVI are frequently employed due to their robust empirical correlation with precipitation; however, their effectiveness diminishes significantly during dormant periods, limiting their applicability in temperate zones and colder seasons [17]. In contrast, elevation is a stable predictor and particularly effective in capturing orographic effects in mountainous regions [18].

Clustering techniques have increasingly been integrated into disaggregation frameworks to enhance spatial coherence and regional model calibration. Methods such as hierarchical clustering [19], K-means clustering [20], and Poisson cluster models [21, 22] have improved precipitation estimates by identifying homogeneous climatic regions. Nevertheless, these clustering approaches typically remain descriptive unless coupled with robust predictive modelling methods [22].

Advances in machine learning, including artificial neural networks (ANN), random forests (RF), and Extreme Gradient Boosting (XGBoost), have demonstrated superior capabilities in modelling complex, non-linear relationships in precipitation data [23–26]. Yet, such models often neglect spatial dependencies and fail to account for climatic

heterogeneity, limiting their effectiveness across diverse landscapes [27]. Recently developed hybrid frameworks, such as Auto-searched Orographic and Atmospheric effects De-trended Kriging (ASOADeK) [28] and microcanonical multiplicative random cascade (MMRC) combined with K-means clustering [20], show promise but frequently lack integration with state-of-the-art machine learning algorithms, thus inadequately capturing critical environmental variables.

In recent years, deep learning-based approaches have gained prominence in precipitation downscaling and disaggregation tasks due to their ability to model complex spatial-temporal dependencies effectively. Convolutional Neural Networks (CNNs) have demonstrated strong performance by capturing spatial structures and local spatial dependencies inherent in precipitation data [29, 30]. Long Short-Term Memory (LSTM) networks, with their capability to model temporal dynamics, have also successfully improved precipitation estimates, particularly in capturing sequence-based dependencies and temporal autocorrelations in rainfall patterns [31]. Moreover, hybrid deep learning frameworks combining CNNs and LSTM have shown enhanced capabilities by simultaneously modelling spatial and temporal features, proving beneficial in regions with complex precipitation regimes [32]. Despite their strong predictive capabilities, deep learning methods often require substantial computational resources and large, high-quality datasets for training. Additionally, they are sensitive to biases in input data, which can propagate through the model and affect performance. These factors may limit their operational applicability, particularly in regions with limited historical observations or inadequate computational infrastructure. Moreover, recent comparative evaluations have demonstrated that machine learning models such as XGBoost not only outperform deep learning approaches like Recurrent Neural Network-Long Short-Term Memory (RNN-LSTM) in terms of predictive accuracy, computational efficiency, and robustness for precipitation estimation [33], but also exhibit superior performance in storm prediction tasks compared to LSTM-based models [34]. These findings reinforce the rationale for adopting XGBoost in our proposed framework.

To address these limitations, this study proposes a novel hybrid modelling approach that integrates Fuzzy Spatial-Temporal Multivariate Clustering with Extreme Gradient Boosting (XGBoost) to effectively disaggregate IMERG satellite precipitation data. Unlike traditional methods, this approach explicitly manages spatial heterogeneity and temporal variability, leveraging fuzzy clustering to capture transitional climatic zones and spatial uncertainty. Coupled with the predictive capabilities of XGBoost, this framework significantly enhances precipitation downscaling accuracy, especially in regions with intricate topography and sparse observational networks.

The specific objectives of this research are:

1. To develop an advanced hybrid modelling approach integrating XGBoost and fuzzy spatial-temporal clustering for improved high-resolution precipitation disaggregation.
2. To reduce satellite precipitation biases by incorporating elevation and ground station observations into the disaggregation process.
3. To rigorously validate model performance across diverse terrains within the Czech Republic, demonstrating robustness and generalizability.

4. To generate a comprehensive, high-resolution precipitation dataset suitable for hydrological and climatological applications, particularly in regions characterised by sparse ground-monitoring networks.

This integrated framework promises substantial improvements for hydrological modelling, water resource management, flood forecasting, drought monitoring, agricultural planning, and climate resilience strategies, particularly in data-limited regions.

Study area

The Czech Republic, located in Central Europe, occupies a strategically significant position at the intersection of several major European river basins. It forms part of the continental divide separating the North, Baltic, and Black Sea drainage systems, making it a crucial hydrological transition zone in Europe. The country hosts approximately 15,536 km of significant watercourses and is characterised by a well-developed river network that includes key rivers such as the Vltava, Elbe (Labe), Morava, and Oder. These rivers serve as vital sources for domestic and transboundary water management and exhibit considerable variability in flow regimes due to climatic and topographic differences.

Climatically, the Czech Republic lies in a temperate continental zone, with significant regional variation due to its diverse topography. Uplands and basins dominate the western and central regions, while the eastern part transitions into lowlands and foothills of the Carpathian Mountains. The Bohemian Massif in the west and the Moravian Highlands in the east create pronounced orographic effects that influence precipitation patterns and runoff behaviour. Elevation ranges from approximately 115 to 1,603 ms above sea level, contributing to nationwide microclimates.

The mean annual precipitation is about 684 mm/year, with a distinct seasonal distribution. Rainfall is generally abundant during spring and summer due to convective storms and cyclonic activity, while winter is marked by lower precipitation and frequent snowfall, particularly in higher altitudes. There is also a regional gradient: the eastern part of the country receives more precipitation during the summer months, whereas the western region tends to receive higher winter precipitation. These seasonal and spatial variations in rainfall directly impact soil moisture, evapotranspiration, and river discharge patterns-key factors in hydrological modeling and flood risk assessment.

The Czech Republic was chosen as the study area for this research due to its hydrological complexity and contrasting climatic regimes, which present an ideal testbed for evaluating high-resolution precipitation disaggregation techniques. The country's diverse landscape, ranging from mountainous terrains to lowland basins and humid and semi-humid zones, allows for robust testing of spatial-temporal models under varying environmental conditions. Moreover, the availability of dense ground-based meteorological and hydrological observational networks supports rigorous validation of satellite-derived precipitation products.

Figure 1b and c present the sub-basin stream order [35] and mean monthly rainfall from IMERG satellite data, respectively, over the Czech Republic during the study period (June 2000–September 2021), illustrating the country's pronounced seasonal rainfall

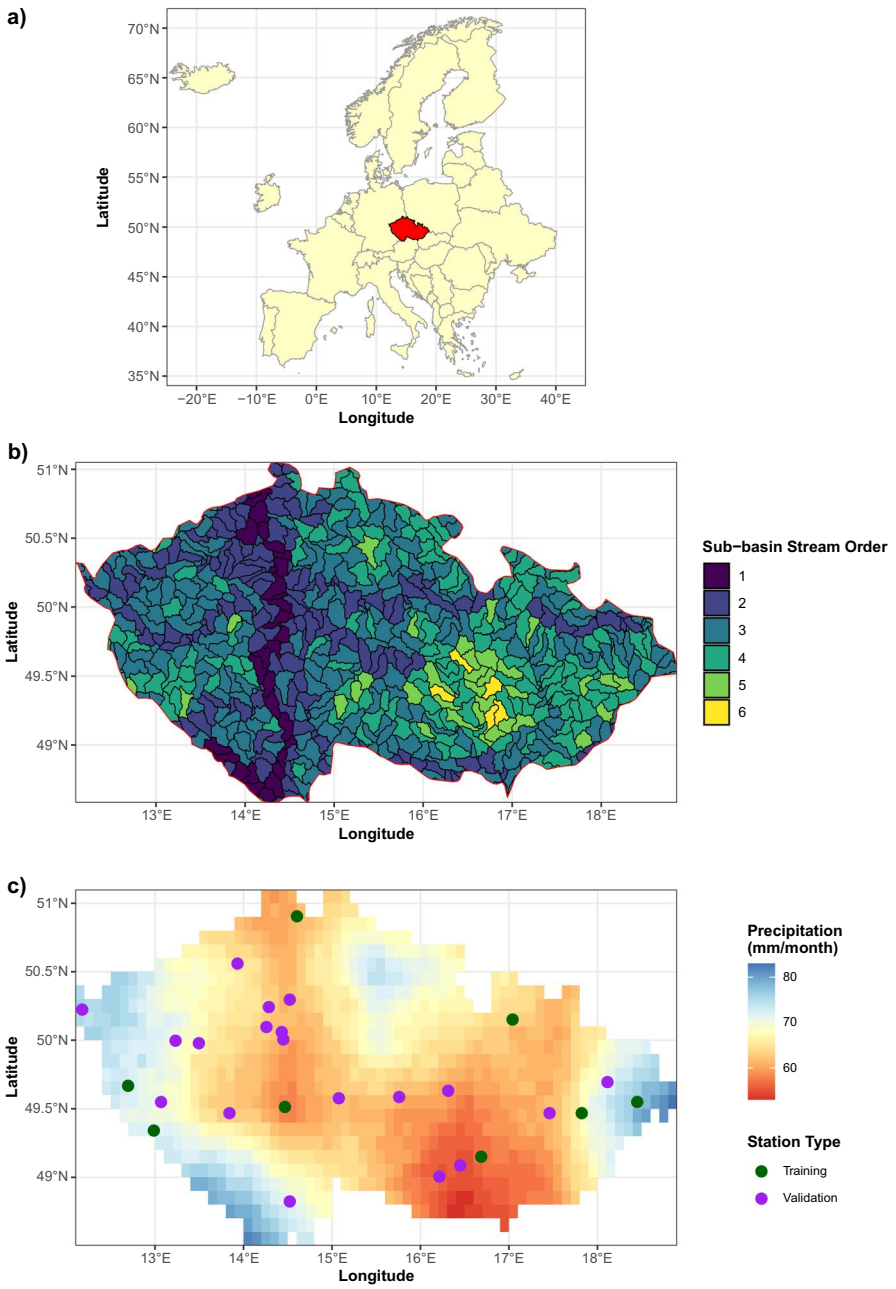

Fig. 1 The study area map is presented in three panels: **a** Europe, with the Czech Republic highlighted in red; **b** sub-basin stream order within the Czech Republic; and **c** the average monthly precipitation over the Czech Republic from June 2000 to September 2021. The precipitation data is derived from IMERG satellite observations, with an original spatial resolution of approximately ~11 km. Training and validation stations are also shown in panel (c) as solid circles, where dark green indicates training stations and purple indicates validation stations

variability and spatial heterogeneity. This variability highlights the need for high-resolution disaggregation methods to capture better-localised precipitation patterns essential for hydrological and climate impact assessments.

Data

Satellite data

The disaggregation accuracy depends on the quality and reliability of input data [36]. The bias-corrected satellite precipitation observations from IMERG, which have been adjusted using gauge data to reduce systematic errors, are used in this study. These corrected estimates are considered less biased compared to raw IMERG satellite products and are, therefore, more suitable for hydrological applications. IMERG is a global precipitation product with 0.1 degrees spatial resolution and available at half-hourly, daily, and monthly temporal resolution. The IMERGV06 final run (gauge corrected) at a monthly scale is employed in this study. This product has high accuracy with respect to the ground measured station [37] and has been available since June 2000. Precipitation patterns have been observed to be linked with elevation directly, as detailed in the study by [38]. In light of this connection, we utilised elevation data in our analysis.

The study utilises the Shuttle Radar Topography Mission (SRTM) Digital Elevation Data Version 4, available online at [39]. This dataset, available from 11th to 22nd February 2002, offers high-quality, global-scale data with a 90-meter spatial resolution offering high-quality, global-scale elevation data.

In situ data

The in-situ data from twenty-seven rain-gauge locations distributed throughout the Czech Republic (Fig. 4a) are used to validate the disaggregated precipitation datasets. This study assumes spatial homogeneity of precipitation within a one-kilometre resolution. The disaggregated precipitation estimates are validated against observations from twenty-seven ground stations, using a point-to-pixel comparison over the period 2000-2021 [40–42].

Methodology

Figure 2 presents the flowchart illustrating the methodology adopted for precipitation disaggregation. Initially, the IMERG satellite precipitation data and the SRTM elevation data were resampled to a common spatial resolution of 1 km to ensure proper spatial alignment for subsequent analysis. The resample() function from the terra R package [43] was utilised to apply three advanced resampling techniques: Cubic, Cubic Spline, and Lanczos.

Despite its widespread use for large raster datasets, nearest-neighbour resampling has known limitations, such as blocky or jagged raster appearances and the potential for pixel value duplication or loss [19, 44, 45]. To overcome these drawbacks and preserve spatial fidelity, interpolation methods that maintain local heterogeneity, edge sharpness, and spatial gradients were prioritised.

The selected resampling techniques offered complementary strengths, described as follows:

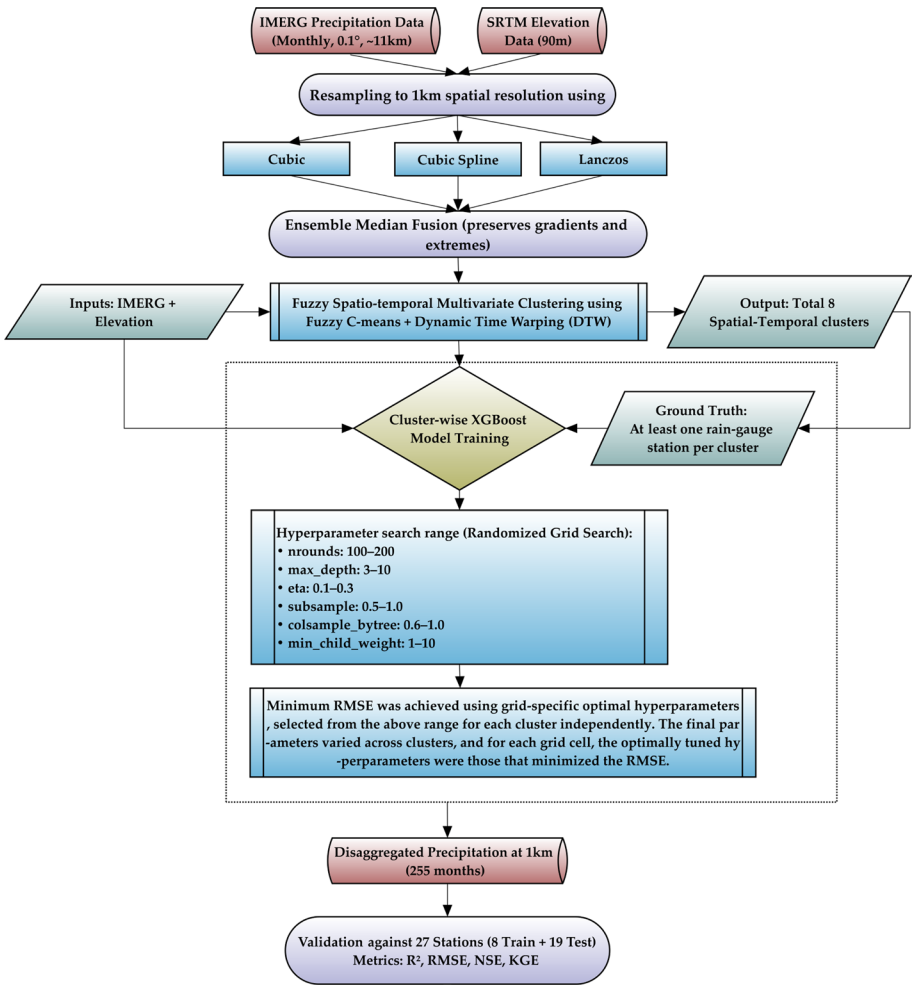

Fig. 2 Flow chart of the methodology for disaggregation of precipitation

Cubic: Pixel values were estimated using the weighted average of the 16 nearest pixels, offering a balance between smooth transitions and edge preservation, thus suitable for continuous variables like precipitation [46].

Cubic Spline: Piecewise cubic polynomials were fitted between data points to achieve smooth continuity, making this method ideal for SRTM elevation data where realistic gradients are essential [45, 47].

Lanczos: Convolution-based resampling using a windowed sinc function was employed to retain fine-scale features and minimise aliasing [48].

Rather than relying on a single method, an ensemble median fusion strategy was implemented. This approach was selected over alternatives like mean or weighted composites due to its robustness to outliers and ability to preserve localised extremes. The median fusion technique effectively synthesised the strengths of all three methods-retaining edge sharpness from Lanczos, smooth gradients from Cubic Spline, and intermediate smoothing from Cubic.

The ensemble-resampled IMERG and SRTM datasets served as foundational inputs for a clustering-based disaggregation pipeline. Fuzzy spatial-temporal multivariate

clustering was then applied to these datasets, using the Fuzzy C-Means (FCM) algorithm [49]. Temporal dissimilarities between time series were quantified using Dynamic Time Warping (DTW), as implemented in the dtwclust R package [50]. This integration enhanced spatial coherence and improved the reliability of disaggregated precipitation.

A total of eight clusters were defined across the Czech Republic, informed by a combination of prior knowledge, domain expertise, heuristic reasoning, and iterative experimentation. The decision to use eight clusters was based on empirical evaluation using the silhouette index-an established internal validation metric-which minimized bias and identified eight as the optimal number of clusters among configurations ranging from 4 to 12. This clustering scheme effectively captures the geographical and climatic heterogeneity of the region, ensuring meaningful spatial representation across distinct hydrological zones. Each cluster includes at least one ground observation station, enhancing the robustness of model training and validation. Time series patterns (Fig. 3) and spatial cluster maps (Fig. 4a) confirm the comprehensive representation of precipitation variability.

To validate cluster quality, the Silhouette index-an internal Cluster Validity Index (CVI)-was computed using the cvi() function in dtwclust [51]. Cluster counts ranging from 4 to 12 were evaluated, and the optimal number of clusters was determined to be eight, based on the highest Silhouette score. This score, ranging from −1 to 1, confirmed strong intra-cluster similarity and inter-cluster separation.

Although each cluster included only one training station, the XGBoost algorithm was selected over simpler alternatives such as LSTM due to its superior ability to model non-linear relationships among elevation, spatial coordinates, and temporal variations. Comparative tests demonstrated that simpler models resulted in higher prediction errors and failed to capture orographic complexity. XGBoost, in contrast, consistently demonstrated superior performance across validation metrics-even with limited training data-and is therefore exclusively highlighted in this study.

A total of 27 ground observation stations were used for training and validation. Within each cluster, the most accurate station (contributing approximately 80% of the data) was used for training, while the remaining 20% was reserved for validation. Additionally, data from 19 independent rain gauge stations were employed to evaluate both the IMERG and the disaggregated outputs.

Both the original IMERG data (at 11 km resolution) and the resampled data (at 1 km resolution via ensemble fusion) were validated against ground observations. This dual-validation strategy enabled assessment of the resampling process's contribution and isolated the improvement resulting from disaggregation.

It must be acknowledged that station-based observations reflect point-scale data, while satellite-derived values represent area-averaged estimates over grid cells. This mismatch can lead to discrepancies, especially in localised extreme events. However, at monthly scales, the effect is generally reduced due to temporal averaging.

The XGBoost model, developed by Chen and Guestrin [52], is based on boosting-a method of combining weak learners to form a strong ensemble. Each decision tree in the ensemble corrects errors from previous iterations, enhancing performance progressively. The xgboost R package was used to implement this method.

For each cluster-specific model, inputs included satellite precipitation, elevation, and station observations. Hyperparameters were tuned via randomized grid search over the following ranges: nrounds = {50, 100, 200}, max_depth = {2, 5, 10}, eta = {0.1, 0.2, 0.3}, gamma = 0, colsample_bytree from 0.5 to 1.0 (in 0.1 increments), min_child_weight = 1–10, and subsample from 0.5 to 1.0 (in 0.05 increments). A repeated 10-fold cross-validation (3 repeats) was conducted, and the configuration minimising RMSE was selected.

Evaluation metrics

The XGBoost model was evaluated using four metrics: Root Mean Square Error (RMSE), Pearson Correlation Coefficient (R^2), Nash–Sutcliffe Efficiency (NSE), and Kling–Gupta Efficiency (KGE).

RMSE: RMSE assessed the agreement between disaggregated and observed precipitation on a pixel basis:

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

where y_i is the observed station value, $\hat{y}_i$ is the disaggregated value, and n is the total number of non-missing grid cells.

R^2: Pearson's correlation coefficient (r) was used to assess the strength of linear association:

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

The coefficient of determination was then calculated as $R^2 = r^2$.

KGE: Kling–Gupta Efficiency was computed as follows [53]:

$$KGE = 1 - \sqrt{(r-1)^2 + \left(\frac{\sigma_{sim}}{\sigma_{obs}} - 1\right)^2 + \left(\frac{\mu_{sim}}{\mu_{obs}} - 1\right)^2} \tag{2}$$

where r is the correlation, σ_{sim} and σ_{obs} are standard deviations, and μ_{sim} and μ_{obs} are means of simulated and observed precipitation, respectively.

NSE: The Nash–Sutcliffe Efficiency [54] was calculated as:

$$NSE = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3}$$

This metric ranges from $-\infty$ to 1, where higher values signify better model performance.

Result

This section shows each cluster's time series patterns of disaggregated precipitation data. Following this, we present the training and validation accuracy of the XGBoost model using designated evaluation metrics. Lastly, we illustrate an evaluation of both IMERG

and the disaggregated data compared with respect to ground-based observational station datasets.

Clustering

The IMERG parent data comprises 1096×255 usable pixels once invalid data or Not-a-Number (NaN) values are excluded. These pixels have a spatial resolution of approximately 11 kms. The number 255 in this calculation shows the raster layers that extend over a monthly timeline from June 2000 to September 2021. A median ensemble resampling technique is applied to the data, enhancing the spatial resolution to one kilometer. As a result, the total number of pixels increased to 119851×255. The analysis has further categorised the data into eight distinct clusters. Each cluster contained a unique set of pixels, the totals of which are computed to be $19848 \times 255, 6157 \times 255, 18077 \times 255, 17491 \times 255, 2848 \times 255, 22953 \times 255, 12757 \times 255,$ and 19684×255 for Clusters 1 through 8 respectively.

A time series plot is created for each raster pixel, grouped by cluster, as demonstrated in Fig. 3. When cross-referenced with elevation data in Figs. 3 and 4 parts a, b, and c, the analysis indicates a pattern of higher precipitation rates in regions of the Czech Republic influenced by elevations. Table 1 presents the range of IMERG precipitation, disaggregated precipitation, accuracy metrics (i.e., R^2, RMSE, NSE, and KGE), and elevation for each cluster. The wide range of IMERG values across clusters, combined with lower accuracy, suggests limited sensitivity of the original satellite product to elevation-induced variability. In contrast, the disaggregated precipitation shows substantial differentiation among clusters, particularly at higher elevations, demonstrating the added value of the clustering-based disaggregation approach. However, for some stations, the disaggregated precipitation was unable to fully capture the observed data range. This indicates that the proposed fuzzy clustering method not only improves spatial resolution but also more effectively captures underlying climatic gradients compared to the original coarse-resolution data, as reflected in both the number and quality of the optimal clusters.

XGBoost model training and validation results

Within each cluster, the model uses 80 per cent of the data from the eight stations, characterised by the highest accuracy for training purposes, pertaining to the information from all eight clusters. This high-quality and carefully quality-controlled station data is integrated into the training phase of the model, subsequently facilitating the prediction of plausible precipitation scenarios within the equivalent cluster. The degree of training and validation accuracy at the pre-selected stations is summarised in Table 2 and Fig. 5. The training and validation accuracy at each station reflects the achievement of satisfactory precipitation disaggregation outcomes. However, the validation accuracy at each station is contingent on the reliability of the training data, as depicted by the evaluation metrics R^2, *RMSE* (mm/month), *NSE*, and *KGE*. Generally, the IMERG precipitation data illustrates a diminished bias at lower elevations, while it presents an intensified bias at higher altitudes. Table 2 suggests that the precipitation data from station O1LYSA01 at the peak elevation exhibits a propensity towards higher bias and extreme values.

Table 1 Statistical performance of IMERG and disaggregated precipitation compared to observations from 27 stations across eight clusters

					IMERG				Disaggregated			
Cluster	**Elevation (m)**	**IMERG (mm/month)**	**Observed (mm/month)**	**Disaggregated (mm/month)**	R^2	**RMSE (mm)**	**NSE**	**KGE**	R^2	**RMSE (mm)**	**NSE**	**KGE**
1	163.39–247.39	1.03–283.28	0.00–298.70	10.41–107.34	0.70	29.31	0.30	0.63	0.63	20.42	0.42	0.67
2	774.46–774.46	0.60–220.23	0.00–276.40	18.00–176.08	0.79	20.65	0.66	0.77	0.80	18.40	0.77	0.86
3	509.07–589.18	1.12–353.67	0.00–323.80	9.14–301.55	0.74	24.08	0.57	0.74	0.74	20.66	0.70	0.80
4	290.98–343.69	1.23–251.44	0.50–291.00	12.24–255.22	0.74	25.03	0.47	0.71	0.74	18.71	0.67	0.81
5	1205.99–1205.99	1.24–330.09	0.90–586.20	29.41–477.21	0.82	57.89	-0.50	0.12	0.85	32.41	0.82	0.89
6	437.32–485.23	1.20–257.18	0.00–196.40	11.83–181.93	0.64	31.64	0.22	0.62	0.65	22.87	0.49	0.74
7	613.36–710.53	0.77–211.45	0.50–200.80	13.80–156.94	0.75	25.32	0.53	0.72	0.75	17.43	0.69	0.83
8	359.85–374.64	1.26–241.71	0.50–291.50	12.01–271.77	0.64	27.92	0.36	0.69	0.65	27.77	0.39	0.69

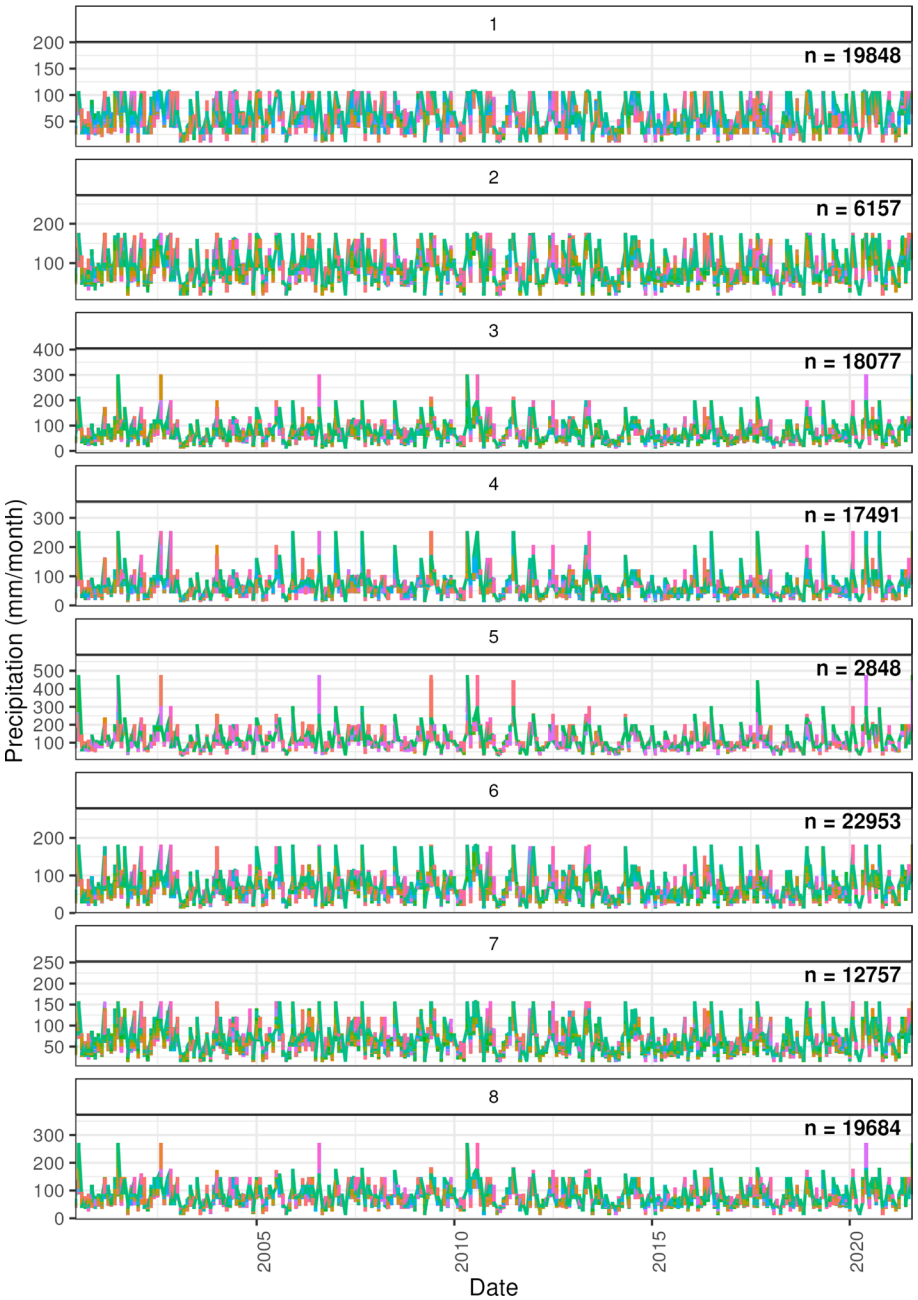

Fig. 3 Fuzzy spatial-temporal cluster time series plot of disaggregated precipitation data. Each line represents the monthly time series of a grid pixel within a specific cluster, with 'n' denoting the total number of unique pixels (based on unique longitude-latitude pairs) in that cluster. Colours represent different individual grid points, each corresponding to a unique longitude-latitude pair within the cluster

To further improve the generalizability of the XGBoost model across varied environmental conditions, we adopted a robust training and validation strategy. Although the number of stations is limited, the dataset includes observations from diverse climatic and topographic zones across the Czech Republic. The 80/20 split of training and validation was complemented with independent testing on 19 additional stations,

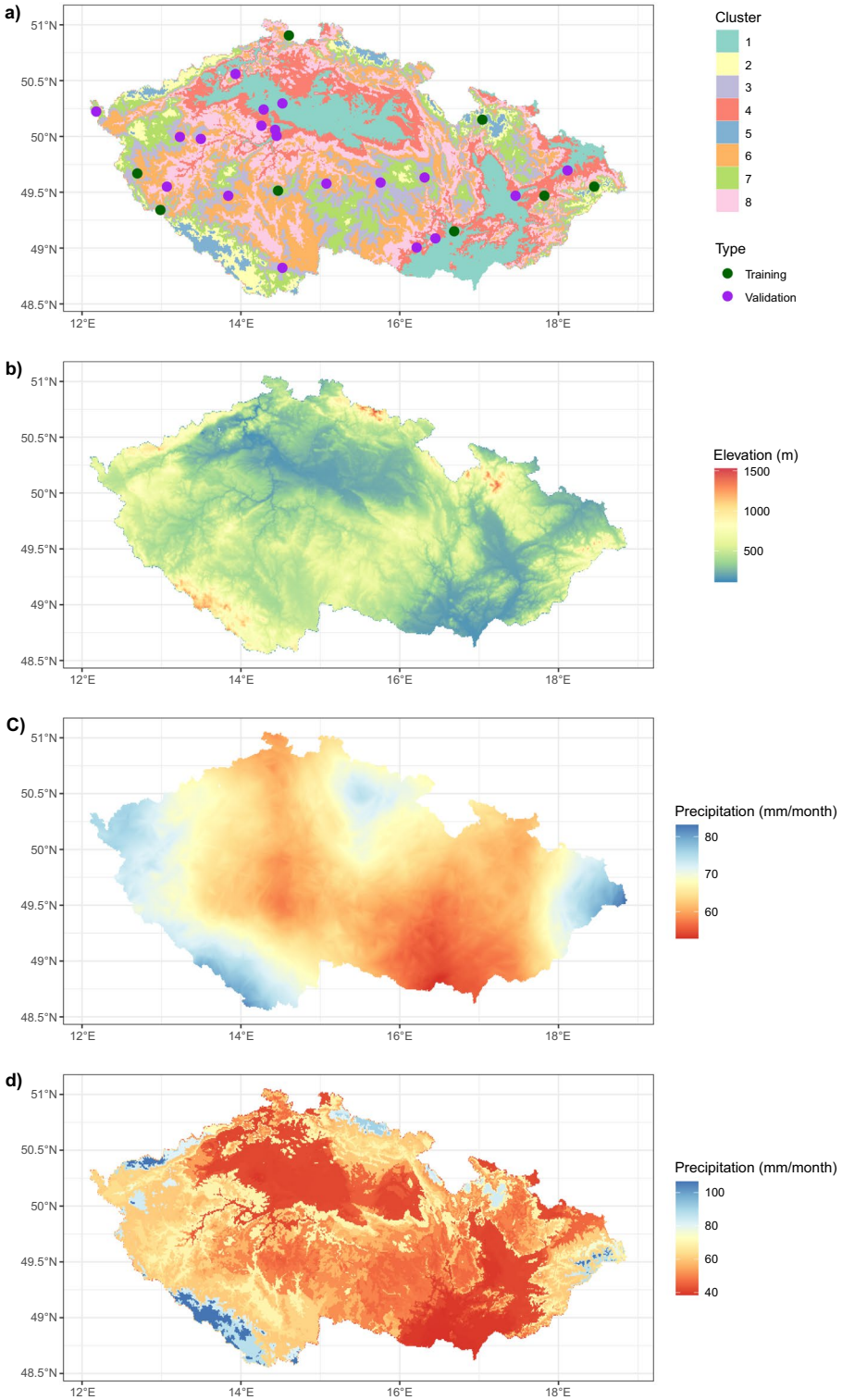

Fig. 4 Spatial visualisation of clustering, elevation, and precipitation over the Czech Republic at 1 km resolution. **a** Spatial distribution of the eight fuzzy spatial-temporal clusters with marked locations of training and validation stations; **b** Topographic elevation map derived from SRTM data, capturing significant terrain variations across the country; **c** Mean monthly precipitation from IMERG satellite data after ensemble resampling; **d** High-resolution disaggregated precipitation map capturing intricate spatial gradients and distinct orographic precipitation patterns, which remain obscured in the original IMERG dataset due to its coarser resolution

Table 2 Accuracy assessment of the selected eight stations using the XGBoost model with respect to the training and validation period against the station measured data

			Training				Validation			
Cluster	**Station**	**Elevation**	R^2	**RMSE (mm/month)**	**NSE**	**KGE**	R^2	**RMSE (mm/month)**	**NSE**	**KGE**
1	B2BTUR01	231.98	0.78	14.21	0.69	0.79	0.69	16.60	0.56	0.75
2	O2BRAN01	774.46	0.81	18.49	0.76	0.84	0.80	18.05	0.79	0.88
3	C2NADV01	589.18	0.93	10.50	0.92	0.91	0.87	12.16	0.86	0.93
4	O3KELC01	308.43	0.88	14.61	0.84	0.86	0.74	22.19	0.72	0.85
5	O1LYSA01	1205.99	0.85	31.36	0.82	0.85	0.85	36.30	0.85	0.91
6	L1VSER01	437.32	0.85	14.45	0.81	0.86	0.76	18.16	0.71	0.84
7	L2PRIM01	613.36	0.88	11.92	0.86	0.88	0.87	13.71	0.82	0.85
8	U2VARN01	374.64	0.90	12.65	0.88	0.88	0.87	16.17	0.76	0.75

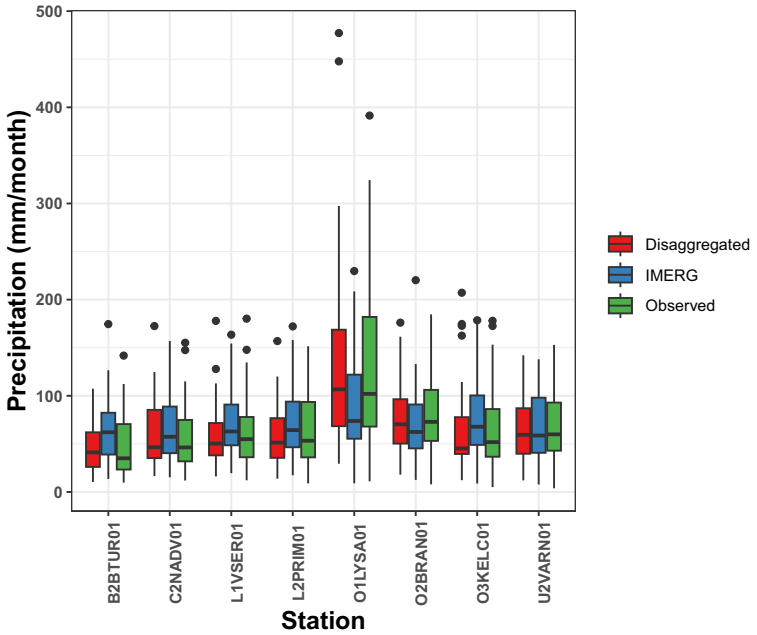

Fig. 5 Comparison of observed, IMERG, and disaggregated precipitation at training stations. Boxplots represent monthly precipitation values at eight selected training stations. Disaggregated values show improved agreement with ground observations compared to the original IMERG data, demonstrating the hybrid model's capacity to reduce bias and enhance accuracy

ensuring spatial diversity. Evaluation metrics including R^2, RMSE, NSE, and KGE showed consistent performance across clusters, with an average R^2 of 0.81 and RMSE of 24.6 mm across validation sites. To benchmark model performance, we compared XGBoost with multiple baselines, including Long Short-Term Memory (LSTM). XGBoost consistently outperformed these alternatives by yielding lower RMSE and higher R^2 values; therefore, we present only the results of XGBoost in the manuscript. To minimize overfitting, we applied ten-fold cross-validation, early stopping, and grid search-based hyperparameter tuning. Furthermore, the residual analysis revealed a slightly higher prediction bias in clusters with greater elevation, such as Cluster 5,

which aligns with the known difficulty of modelling orographic effects. However, the integration of elevation and cluster ID as features mitigated much of this bias. The ensemble-based learning structure of XGBoost, paired with multivariate fuzzy clustering, thus provides a resilient framework for disaggregating precipitation across heterogeneous landscapes.

Assessment of IMERG and disaggregated data with respect to ground observed station datasets

Our findings, which are shown in detail in Table 3, indicate that there is a high correlation and low error values (*e.g.*, $highR^2$, $lowRMSE$) between the majority of the precipitation data collected from the station and the data produced by the IMERG. This conclusion is further substantiated by the evaluation metrics used for this study. These metrics are reliable tools for comparing and validating the consistency and accuracy of both data sets, confirming their close correlation. The key performance

Table 3 Accuracy assessment of the disaggregated and IMERG precipitation against the 27 stations measured data for 255 months

			Disaggregataed precipitation				IMERG precipitation			
Cluster	Station	Elevation	R^2	RMSE (mm/month)	NSE	KGE	R^2	RMSE (mm/month)	NSE	KGE
1	P1KRAL01	239.92	0.59	20.09	0.40	0.70	0.67	30.74	0.16	0.59
1	P1PKAR01	238.70	0.63	19.12	0.44	0.71	0.68	32.80	0.04	0.53
1	P2TUHA01	163.39	0.63	19.07	0.47	0.72	0.73	28.66	0.29	0.62
1	B2BTUR01	231.98	0.76	14.72	0.67	0.78	0.72	23.97	0.42	0.68
1	O3PRER01	218.76	0.68	18.95	0.47	0.66	0.74	25.95	0.41	0.68
1	O1MOSN01	247.39	0.58	28.15	0.09	0.48	0.70	32.61	0.37	0.67
2	O2BRAN01	774.46	0.80	18.40	0.77	0.86	0.79	20.65	0.66	0.77
3	C1KOCE01	510.28	0.79	20.26	0.70	0.73	0.81	23.86	0.54	0.69
3	C2NADV01	589.18	0.92	10.85	0.91	0.92	0.90	14.56	0.83	0.83
3	P3NETV01	572.53	0.59	33.12	0.35	0.57	0.59	38.01	0.14	0.52
3	P3KOSE01	509.07	0.81	16.74	0.80	0.87	0.83	17.86	0.74	0.82
3	P3PRIB01	521.23	0.84	15.11	0.81	0.89	0.83	18.80	0.71	0.80
4	P1PLIB01	290.98	0.72	19.27	0.66	0.82	0.75	25.74	0.41	0.68
4	B2DZBA01	343.69	0.64	20.13	0.51	0.74	0.66	27.16	0.30	0.65
4	B2MBRA01	306.53	0.67	18.82	0.53	0.74	0.70	23.67	0.44	0.70
4	O3KELC01	308.43	0.84	16.41	0.81	0.88	0.79	23.35	0.62	0.76
5	O1LYSA01	1205.99	0.85	32.41	0.82	0.89	0.82	57.89	-0.50	0.12
6	L1VSER01	437.32	0.83	15.27	0.79	0.86	0.80	20.69	0.69	0.81
6	L2MANE01	485.23	0.62	25.30	0.35	0.64	0.61	35.68	-0.01	0.51
6	L2KRAL01	444.00	0.59	26.34	0.27	0.61	0.61	36.02	-0.06	0.49
7	L3AS0001	666.19	0.74	18.56	0.71	0.84	0.76	27.46	0.51	0.69
7	L2PRIM01	613.36	0.88	12.30	0.85	0.88	0.85	19.23	0.73	0.81
7	U1MILE01	663.96	0.65	20.58	0.49	0.74	0.65	30.41	0.24	0.61
7	B2NEDV01	710.53	0.76	17.19	0.61	0.72	0.75	22.62	0.54	0.74
8	L1STAN01	359.85	0.65	34.37	0.15	0.56	0.65	33.51	0.12	0.57
8	P1PRUZ01	371.17	0.66	30.84	0.12	0.60	0.65	31.30	0.13	0.59
8	U2VARN01	374.64	0.89	13.43	0.86	0.86	0.85	15.36	0.81	0.85

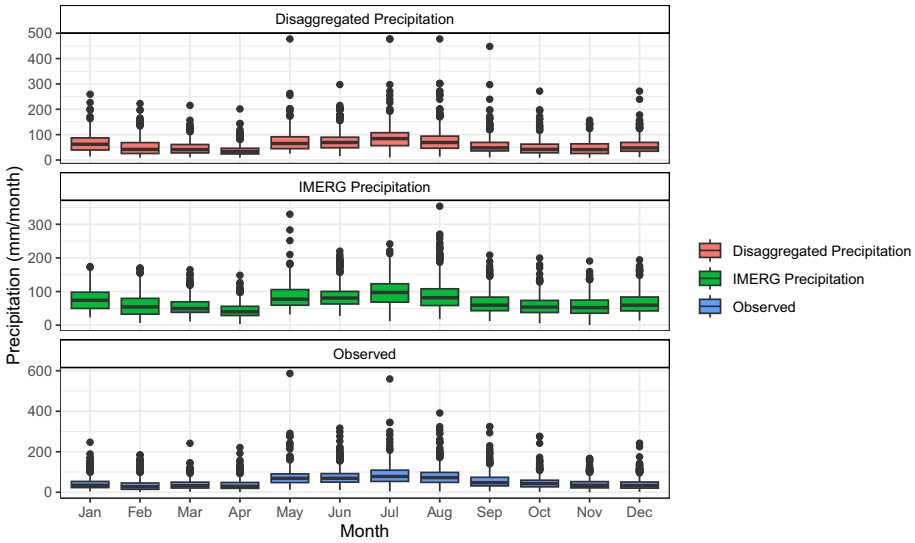

Fig. 6 Monthly precipitation comparison across all stations. Boxplots display observed, IMERG, and disaggregated precipitation distributions from January to December. The disaggregated outputs more closely align with observed values, particularly in summer months when precipitation variability is highest

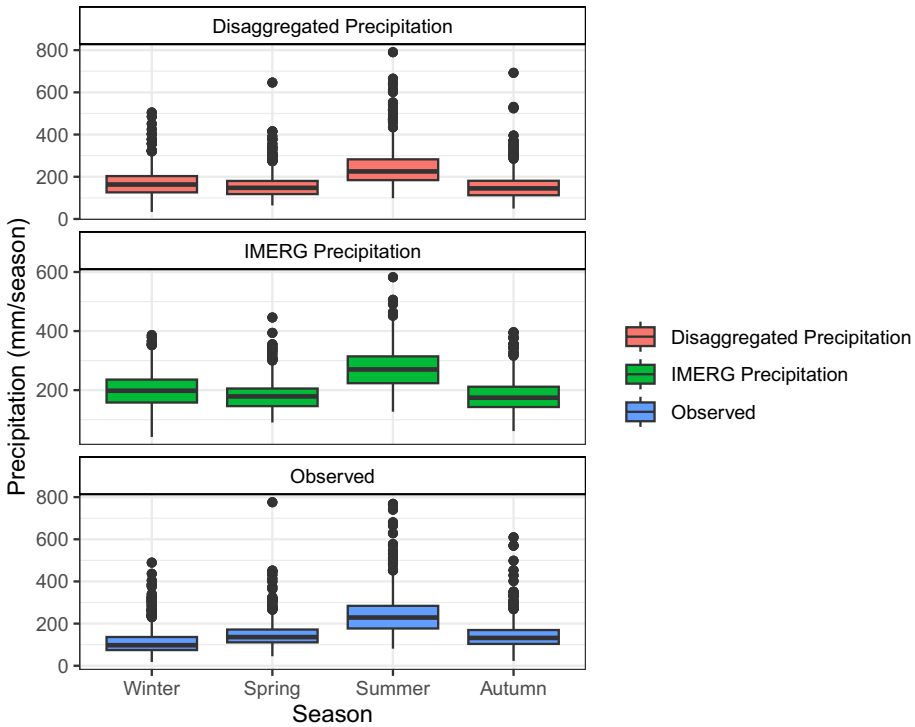

Fig. 7 Seasonal-scale comparison of precipitation. Precipitation aggregated by meteorological seasons (Winter, Spring, Summer, Autumn) across all stations. Disaggregated precipitation effectively captures seasonal trends and extreme values, outperforming the coarse IMERG product

indicators, specifically the R^2, the RMSE (mm/month) NSE, and the KGE, for the IMERG data across all the stations, range between 0.61—0.90, 14.56—57.79, —0.50—

0.83, and 0.12-85, respectively. Notably, the data from IMERG at higher elevations demonstrates a higher bias compared to those at lower elevations. Moreover, the IMERG data patterns are shown at monthly and seasonal scales with respect to each station by the box plots illustrated in Figs. 6 and 7. IMERG is adept at capturing patterns at monthly and seasonal scales, as substantiated by Figs. 6 and 7. The precision of the IMERG data remains satisfactory within an elevation range spanning from 250 to 774 ms. Nevertheless, IMERG data does not adequately represent the heterogeneous spatio-temporal precipitation patterns over the Czech Republic, thereby necessitating further improvements in the accuracy of IMERG data, potentially through higher resolution data.

IMERG precipitation data is enhanced through disaggregation, demonstrating greater congruency with ground-observed station data, as indicated in Table 3. The comprehensive accuracy spanning all stations, including training and validation phases, is also depicted in Table 3. Additionally, the precision of disaggregated precipitation surpasses that of IMERG at high altitudes, notably at the O1LYSA01 station, as depicted in Figs. 6 and 7, and Table 3. A diminished bias in IMERG data results in improved accuracy of disaggregated precipitation, a phenomenon demonstrated by Tables 2 and 3. The reliability of IMERG data significantly influences the precision of disaggregated precipitation, with elevated accuracy observed in regions where IMERG precipitation data is deemed satisfactory. Disaggregation aids in curtailing the bias of IMERG by enhancing the spatial resolution to 1 km, thereby offering a more comprehensive representation of spatial variability.

Improvements in the accuracy of IMERG data are achieved through station-specific disaggregation, which reduces bias and better captures extreme precipitation events, as shown in Fig. 8. Evaluation metrics-R^2, RMSE (mm/month), NSE, and KGE-summarised in Table 3, support the performance gains from disaggregation. Figures 4c and d illustrate the enhanced spatio-temporal variability of precipitation captured in the disaggregated data, which is not as evident in the original IMERG product.

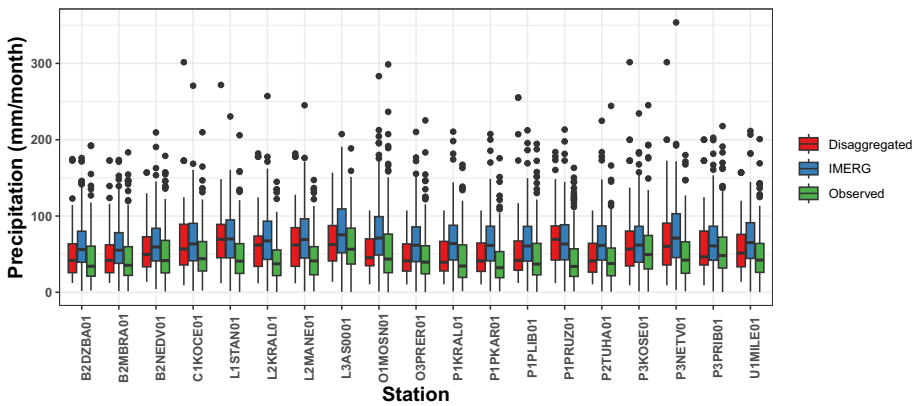

Fig. 8 Validation results across independent stations. Boxplots for 19 independent validation stations demonstrate enhanced spatial and statistical agreement between disaggregated precipitation and observed data. The proposed hybrid model exhibits greater consistency and reliability than the IMERG dataset, especially in topographically complex areas

Despite the overall improvements, model performance varies across stations, especially at higher elevations. For example, station O1LYSA01, located at 1205.99 m, shows relatively high RMSE and negative NSE values. A broader evaluation of model performance reveals a trend of reduced accuracy at elevations above approximately 800 m, indicating a potential influence of terrain complexity. These spatial differences in performance metrics suggest that disaggregation improves overall accuracy, but localised discrepancies remain, particularly in complex topographic settings.

The disaggregated precipitation map (Fig. 4d) exhibits distinct spatial features-particularly around 50° N and 14°E-that may appear as linear or stream-like bands. However, a comparison with the resampled IMERG data (Fig. 4c) and the elevation map (Fig. 4b) indicates that these patterns are not resampling artefacts but rather reflect the influence of orographic effects on precipitation distribution. The ensemble resampling approach, which incorporates cubic, spline, and Lanczos methods, was specifically designed to preserve localised gradients. This is evident in the spatial consistency between the resampled IMERG (Fig. 4c) and the original IMERG precipitation distribution (Fig. 1c), both of which align closely with terrain features. Given the strong control of elevation and topographic barriers-especially in the western and northern sub-basins-on regional precipitation, the enhanced spatial resolution of the disaggregated product more clearly captures these dependencies. While some cluster boundaries in Fig. 1c may appear abrupt, this is primarily attributed to sharp elevation gradients and the associated orographic influences, which are faithfully preserved through the ensemble resampling process. Nonetheless, future refinements could involve terrain-adaptive smoothing or additional filtering to further validate these spatial patterns against station-level observations.

Monthly and seasonal patterns of disaggregated precipitation also align more closely with observed data compared to IMERG, as shown in Figs. 6 and 7. These figures further demonstrate the improved representation of precipitation variability across time scales in the disaggregated product.

Discussion

Integrated multi-satellite hydro-meteorological datasets, such as IMERG precipitation products, are essential ultra-big data resources for global-scale weather and climate analyses. However, their application at national or regional levels requires high-resolution modelling to accurately capture local-scale variability. While approaches such as exponential regression, MLR, and GWR have been widely used for disaggregation, their limitations-including stationary assumptions and multicollinearity-are exacerbated by a lack of integration with spatial-temporal clustering and elevation sensitivity.

To address these gaps, we propose a novel hybrid modelling approach that synergistically integrates fuzzy multivariate spatio-temporal clustering with the XGBoost algorithm. This framework dynamically learns non-linear spatial and temporal dependencies within data clusters sharing similar climatic and topographic features. Fuzzy clustering enhances spatial and temporal coherence, while XGBoost effectively models variable interactions, thereby reducing within-cluster variability and bias-particularly in high-elevation areas. Benchmarking results demonstrate that the hybrid model consistently

outperforms alternatives across all evaluation metrics (R^2, RMSE, NSE, and KGE), affirming its robustness and generalizability in hydro-meteorological applications.

Recent developments in statistical downscaling, such as the En-GARD framework proposed by Gutmann et al. [55], apply analog-based methods that integrate empirical quantile mapping with bias correction. These methods perform well in preserving precipitation characteristics, particularly in complex terrains. However, analog approaches typically require extensive, high-quality historical data and may underperform in data-scarce or non-stationary environments.

By contrast, our method effectively captures non-linear interactions among predictors such as elevation, spatial coordinates, and temporal variability, without the need for long-term historical records. Moreover, compared to deep learning models like LSTM-which require substantial computational resources and large training datasets-XGBoost offers superior interpretability and delivers comparable or even improved performance relative to both LSTM [34] and RNN-LSTM models [33]. These characteristics make the proposed method particularly well-suited for operational applications in data-scarce environments.

Applied over the Czech Republic, our approach downscales IMERG precipitation to a finer spatial resolution by integrating elevation data within the clustering framework. The IMERG dataset, calibrated with rain gauge data [56], generally aligns with ground observations, although greater discrepancies occur at higher altitudes. For example, at station O1LYSA01 (1205 m), pronounced bias was observed, consistent with the variability of precipitation over short distances at high elevations [57], and the coarse spatial resolution (11×11 km) of IMERG pixels [58].

To mitigate this, we combined elevation-based clustering with XGBoost, enabling finer-scale disaggregation. Model validation compared disaggregated and resampled IMERG estimates with both ground station and cluster-level data. Sources of error-such as spatial heterogeneity, grid-cell mismatch [59], and wind-induced gauge undercatch [60]-can lead to elevated bias, particularly in mountainous regions. By incorporating elevation and precipitation into clustering, the model better captures spatial variability, which is critical for realistic environmental and climate modelling [61–63].

Clustering based on spatiotemporal precipitation and climate characteristics, followed by training using a representative high-quality station per cluster, improves intra-cluster variability representation. Including elevation as a predictor in XGBoost further reduces performance variability, particularly in complex terrain. This ensures the predictions are grounded in the most reliable observational data for each region.

Eight stations were selected for model training based on data quality. While domain-informed clustering effectively captured broad spatial patterns, some heterogeneity-especially in mountainous zones-remains unresolved. Validation performance was slightly lower than training, with R^2 values ranging from 0.59 to 0.87 versus 0.85 to 0.93 in training. Notably, RMSE increased at high-elevation stations like O1LYSA01, rising from 31.36 mm/month to 36.30 mm/month. Although disaggregated precipitation showed better agreement with observations, NSE values varied, reflecting spatial performance variability.

These discrepancies are expected, due to limitations in IMERG's resolution, spatial mismatches between pixels and gauges, and undercatch effects [60]. Elevation and

cluster ID helped mitigate some issues in the XGBoost model, but residual uncertainty remains. Nevertheless, ten-fold cross-validation and hyperparameter tuning (e.g., nrounds, maxdepth, eta) improved model generalisation and reduced overfitting.

Model reliability strongly depends on the quality of both IMERG and ground station data. The use of gauge-corrected IMERG [56] offered a strong baseline. Disaggregation significantly improved spatial detail and accuracy. At station O1LYSA01, RMSE dropped from 57.89 mm/month (IMERG) to 32.41 mm/month (disaggregated), while NSE improved from −0.50 to 0.82 (Table 3), highlighting effective bias mitigation. However, residual bias persists due to unresolved spatial mismatches and orographic complexity, although the interpretability and computational efficiency of disaggregated precipitation make it well-suited for operational deployment.

Broadly, the high-resolution disaggregated precipitation produced by this method can substantially enhance hydrological forecasting, particularly for flash flood prediction, reservoir management, and drought assessment. This spatial precision also supports data-driven water policy planning in topographically complex or data-scarce regions, thus contributing to resilient climate adaptation strategies.

Limitations and future work

While this study advances the disaggregation of IMERG satellite precipitation data through a hybrid approach combining multivariate clustering and XGBoost, several limitations must be acknowledged. First, model training utilised only eight stations, selected based on the availability of daily-scale observational data. As a result, the clustering framework-guided by expert judgment rather than fully objective data-driven methods-may not adequately reflect the complex and heterogeneous precipitation dynamics across mountainous terrain.

Second, validation performance revealed signs of overfitting, particularly at higher elevations, where there was a noticeable decline in accuracy from training to validation phases. This suggests that the model may not generalise well in areas where topographic and meteorological conditions deviate significantly from the training data.

Third, uncertainties in precipitation measurements at high-altitude stations-primarily due to undercatch caused by strong winds-can propagate into the disaggregated outputs, impacting model reliability. Moreover, inherent biases in the IMERG product, especially in cold-cloud or snowy environments, may further contribute to inaccuracies in the final results.

A further limitation concerns the spatial generalisation of precipitation estimates in high-elevation zones. Pixels situated in different corners of the domain but sharing similar elevations may yield comparable precipitation values, even when their broader climatic contexts differ. While our fuzzy clustering framework incorporates both elevation and spatiotemporal precipitation patterns, reliance on a single representative station per cluster introduces uncertainty, particularly in geographically remote or topographically complex areas.

Additionally, using elevation as the primary topographic predictor may overlook key localized orographic processes-such as variations between windward and leeward slopes-that strongly influence precipitation. To improve spatial realism and predictive accuracy, especially in mountainous regions, future efforts should incorporate

complementary terrain and meteorological variables (e.g., slope orientation, aspect, prevailing wind direction), alongside a denser and more spatially distributed training network.

Although model hyperparameters were optimised via ten-fold cross-validation, detailed sensitivity analyses were not performed. This omission limits confidence in the robustness of the model under varying climatic and geographic conditions. Furthermore, uncertainties associated with extreme precipitation events were not quantified, which may restrict the applicability of the disaggregated outputs for flood forecasting and extreme weather risk assessments.

Another notable limitation is that the disaggregated precipitation outputs do not fully capture the wide observed range of monthly precipitation, especially in clusters with high rainfall variability. This constraint arises from the dependency of the disaggregated values on a single training station per cluster, which may limit the representation of intra-cluster variability and extreme values. As evident in Table 1, while disaggregation improves upon IMERG estimates, it still underrepresents the full precipitation range observed across the 27 validation stations. Incorporating multiple training stations per cluster and more diverse training data, depending on the availability of reliable historical rain gauge records, could enhance the model's ability to replicate observed extremes.

Future research should therefore prioritise the integration of denser observational datasets, adoption of objective clustering methodologies, implementation of robust sensitivity analyses, and explicit quantification of uncertainties related to extreme events. These enhancements will strengthen the generalizability, accuracy, and operational value of satellite-based precipitation disaggregation approaches across regional and global hydrological applications.

Conclusion

In this study, we developed a hybrid learning approach coupling multivariate clustering with the XGBoost model to spatially disaggregate IMERG satellite precipitation data over the Czech Republic. The key findings and outcomes are summarised numerically below:

1. The hybrid method effectively disaggregates coarse-resolution IMERG precipitation data to a finer 1 km resolution by incorporating elevation data (SRTM) and clustering similar precipitation patterns, significantly improving agreement with observed station data.
2. Original IMERG data show limited variability (0.25–384 mm/month), particularly at high elevations. The proposed method substantially expands this range (e.g., 29–477 mm/month at elevations 890–1528 m), better capturing local precipitation variability.
3. IMERG precipitation exhibits considerable biases at higher elevations (e.g., RMSE = 57.89 mm/month, NSE = −0.50 at station O1LYSA01, 1205.99 m). While our method significantly reduces these biases, particularly at lower elevations (250–774 m), it does not entirely eliminate inaccuracies at high altitudes.

4. The developed method enhances both spatial resolution and accuracy, making the resulting dataset valuable for various hydrological and meteorological applications, particularly under climate change scenarios involving frequent droughts and floods.
5. Further research is recommended, incorporating additional environmental variables, elevation-dependent bias corrections, extensive observational datasets, and comprehensive uncertainty quantification, especially for extreme precipitation events. These steps will enhance the applicability and reliability of disaggregated precipitation data for broader global hydrological and extreme weather studies.

In conclusion, our approach represents a significant technical advancement in producing accurate, high-resolution precipitation datasets suitable for robust water resource management and policy-making, particularly in areas with limited observational coverage.

Acknowledgements
Computational resources were supplied by the project "e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic". The authors would like to thank the Czech Hydro-meteorological Institute for providing the ground-observed precipitation.

Author contributions
Ujjwal Singh conceptualised the study, designed the methodology, performed simulations, conducted data analysis and visualisation, and drafted the original manuscript & revisions. Petr Maca and Rama Rao Nidamanuri actively contributed to the design of methodology, data acquisition, scientific discussions, manuscript review, and editing. Sadaf Nasreen, Gaurav Tripathi, Pragya Mehrishi, Rajani Kumar Pradhan, Zuzana Bestakova, Vivek Vikram Singh, K. C. Gouda, Laxmi Kant Sharma, Kiran Jalem, Akhilesh Singh Raghubanshi, Yannis Markonis, Rakovec Oldřich, and Martin Hanel provided critical feedback, participated in discussions, and contributed to the review and revision of the manuscript. Petr Maca and Martin Hanel helped with funding this research.

Funding
This study was funded by the Internal Grant Agency of the Faculty of Environmental Sciences of the Czech University of Life Sciences (Grant No. 2021B0021, 2022B0018 and 2023B0032 to the Corresponding author (Ujjwal Singh). This publication has been produced as part of the project RUR - Region for university, university for region (Reg. No. CZ.10.02.01/00/22_002/0000210), with the financial support of the European Union. This research has been also supported by the Ministry of Education, Youth and Sports of the Czech Republic (grant AdAgriF - Advanced methods of greenhouse gases emission reduction and sequestration in agriculture and forest landscape for climate change mitigation (CZ.02.01.01/00/22_008/0004635).

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Faculty of Environmental Sciences, Czech University of Life Sciences Prague, Kamýcá 129, 165 00 Praha-Suchdol, Czech Republic. [2]Department of Civil Engineering, Indian Institute of Technology Bombay, Powai, Maharashtra 400076, India. [3]Department of Physical Geography and Geoecology, Faculty of Science, Charles University. Albertov 6, 128 00 Prague 2, Czech Republic. [4]Faculty of Forestry and Wood Sciences, Czech University of Life Sciences, Kamýcká 129, 165 00 Praha - Suchdol, Czech Republic. [5]CSIR-Fourth Paradigm Institute, NAL Belur Campus, Bengaluru, India. [6]Department of Environmental Science, Central University of Rajasthan, Ajmer, India. [7]Department of Geoinformatics, Central University of Jharkhand, Ranchi, India. [8]Department of Earth and Space Sciences, Indian Institute of Space Science and Technology, Valiamala, Thiruvananthapuram 695547, India. [9]Institute of Environment and Sustainable Development, Banaras Hindu University, Varanasi, UP 221005, India.

Received: 8 August 2023 Accepted: 22 May 2025
Published online: 20 June 2025

References

1. Tapas MR, Do SK, Etheridge R, Lakshmi V, et al. Investigating the impacts of climate change on hydroclimatic extremes in the Tar-Pamlico River Basin, North Carolina. J Environ Manage. 2024;363:121375.
2. Vu CH, Nguyen BQ, Tran T-N-D, Vo DN, Arshad A. Climate change-driven hydrological shifts in the Kon-Ha Thanh River basin. Water. 2024;16(23):3389.
3. Nguyen BQ, Van Binh D, Tran T-N-D, Kantoush SA, Sumi T. Response of streamflow and sediment variability to cascade dam development and climate change in the Sai Gon Dong Nai River basin. Clim Dyn. 2024;62(8):7997–8017.
4. Tran T-N-D, Nguyen BQ, Grodzka-Łukaszewska M, Sinicyn G, Lakshmi V. The role of reservoirs under the impacts of climate change on the Srepok River basin, central highlands of Vietnam. Front Environ Sci. 2023;11:1304845.
5. Lakshmi V, et al. Enhancing human resilience against climate change: assessment of hydroclimatic extremes and sea level rise impacts on the eastern shore of Virginia, United States. Sci Total Environ. 2024;947:174289.
6. Tapas MR, Etheridge R, Finlay CG, Peralta AL, Bell N, Xu Y, Lakshmi V, et al. A methodological framework for assessing sea level rise impacts on nitrate loading in coastal agricultural watersheds using swat+: a case study of the Tar-Pamlico River basin, North Carolina, USA. Sci Total Environ. 2024;951:175523.
7. Aryal A, Tran T-N-D, Kumar B, Lakshmi V. Evaluation of satellite-derived precipitation products for streamflow simulation of a mountainous Himalayan watershed: a study of Myagdi Khola in Kali Gandaki basin, Nepal. Remote Sens. 2023;15(19):4762.
8. Le M-H, Zhang R, Nguyen BQ, Bolten JD, Lakshmi V, et al. Robustness of gridded precipitation products for Vietnam basins using the comprehensive assessment framework of rainfall. Atmos Res. 2023;293:106923.
9. Xiao S, Xia J, Zou L. Evaluation of multi-satellite precipitation products and their ability in capturing the characteristics of extreme climate events over the Yangtze River basin, China. Water. 2020;12(4):1179.
10. Pradhan RK, Markonis Y, Godoy MRV, Villalba-Pradas A, Andreadis KM, Nikolopoulos EI, Papalexiou SM, Rahim A, Tapiador FJ, Hanel M. Review of gpm imerg performance: a global perspective. Remote Sens Environ. 2022;268:112754.
11. Tran T-N-D, Nguyen BQ, Zhang R, Aryal A, Grodzka-Łukaszewska M, Sinicyn G, Lakshmi V. Quantification of gridded precipitation products for the streamflow simulation on the Mekong River basin using rainfall assessment framework: a case study for the Srepok River subbasin, central highland Vietnam. Remote Sens. 2023;15(4):1030.
12. Tapas MR, Etheridge R, Le M-H, Hinckley B, Nguyen VT, Lakshmi V, et al. Evaluating combinations of rainfall datasets and optimization techniques for improved hydrological predictions using the swat+ model. J Hydrol Reg Stud. 2025;57:102134.
13. Ghorbanpour AK, Hessels T, Moghim S, Afshar A. Comparison and assessment of spatial downscaling methods for enhancing the accuracy of satellite-based precipitation over lake Urmia basin. J Hydrol. 2021;596:126055.
14. Chen S, Zhang L, Zhang Y, Guo M, Liu X. Evaluation of tropical rainfall measuring mission (trmm) satellite precipitation products for drought monitoring over the middle and lower reaches of the Yangtze River basin, China. J Geogr Sci. 2020;30(1):53–67.
15. Chen S, Zhang L, She D, Chen J. Spatial downscaling of tropical rainfall measuring mission (trmm) annual and monthly precipitation data over the middle and lower reaches of the Yangtze River basin, China. Water. 2019;11(3):568.
16. Foody G. Geographical weighting as a further refinement to regression modelling: an example focused on the ndvi-rainfall relationship. Remote Sens Environ. 2003;88(3):283–93.
17. Xu S, Wu C, Wang L, Gonsamo A, Shen Y, Niu Z. A new satellite-based monthly precipitation downscaling algorithm with non-stationary relationship between precipitation and land surface characteristics. Remote Sens Environ. 2015;162:119–40.
18. Zhang T, Li B, Yuan Y, Gao X, Sun Q, Xu L, Jiang Y. Spatial downscaling of trmm precipitation data considering the impacts of macro-geographical factors and local elevation in the three-river headwaters region. Remote Sens Environ. 2018;215:109–27.
19. Lober C, Fayne J, Hashemi H, Smith LC. Bias correction of 20 years of imerg satellite precipitation data over Canada and Alaska. J Hydrol Reg Stud. 2023;47:101386.
20. Deka P, Saha U. Introduction of k-means clustering into random cascade model for disaggregation of rainfall from daily to 1-hour resolution with improved preservation of extreme rainfall. J Hydrol. 2023;620:129478.
21. Koutsoyiannis D, Onof C. Rainfall disaggregation using adjusting procedures on a Poisson cluster model. J Hydrol. 2001;246(1–4):109–22.
22. Segond M-L, Onof C, Wheater H. Spatial-temporal disaggregation of daily rainfall from a generalized linear model. J Hydrol. 2006;331(3–4):674–89.
23. Hsu K-L, Gao X, Sorooshian S, Gupta HV. Precipitation estimation from remotely sensed information using artificial neural networks. J Appl Meteorol. 1997;36(9):1176–90.
24. Vishwakarma BD, Zhang J, Sneeuw N. Downscaling grace total water storage change using partial least squares regression. Sci Data. 2021;8(1):95.
25. Zhu H, Liu H, Zhou Q, Cui A. Towards an accurate and reliable downscaling scheme for high-spatial-resolution precipitation data. Remote Sens. 2023;15(10):2640.
26. Sebbar B-E, Khabba S, Merlin O, Simonneaux V, Hachimi CE, Kharrou MH, Chehbouni A. Machine-learning-based downscaling of hourly era5-land air temperature over mountainous regions. Atmosphere. 2023;14(4):610.
27. Arshad A, Zhang W, Zhang Z, Wang S, Zhang B, Cheema MJM, Shalamzari MJ. Reconstructing high-resolution gridded precipitation data using an improved downscaling approach over the high altitude mountain regions of Upper Indus Basin (uib). Sci Total Environ. 2021;784:147140.
28. Guan H, Wilson JL, Xie H. A cluster-optimizing regression-based approach for precipitation spatial downscaling in mountainous terrain. J Hydrol. 2009;375(3–4):578–88.
29. Wang F, Tian D, Lowe L, Kalin L, Lehrter J. Deep learning for daily precipitation and temperature downscaling. Water Resour Res. 2021;57(4):2020–029308.
30. Li W, Pan B, Xia J, Duan Q. Convolutional neural network-based statistical post-processing of ensemble precipitation forecasts. J Hydrol. 2022;605:127301.

31. Zheng C, Tao Y, Zhang J, Xun L, Li T, Yan Q. Tise-lstm: a lstm model for precipitation nowcasting with temporal interactions and spatial extract blocks. Neurocomputing. 2024;590:127700.
32. Li X, Xu W, Ren M, Jiang Y, Fu G. Hybrid CNN-LSTM models for river flow prediction. Water Supply. 2022;22(5):4902–19.
33. Santoro D, Ciano T, Ferrara M. A comparison between machine and deep learning models on high stationarity data. Sci Rep. 2024;14(1):19409.
34. Frifra A, Maanan M, Maanan M, Rhinane H. Harnessing lstm and xgboost algorithms for storm prediction. Sci Rep. 2024;14(1):11381.
35. Lehner B, Grill G. Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems. Hydrol Process. 2013;27(15):2171–86.
36. Eichhorn S. Disaggregating population data and evaluating the accuracy of modeled high-resolution population distribution-the case study of Germany. Sustainability. 2020;12(10):3976.
37. Sun W, Sun Y, Li X, Wang T, Wang Y, Qiu Q, Deng Z. Evaluation and correction of gpm imerg precipitation products over the capital circle in Northeast China at multiple spatiotemporal scales. Adv Meteorol. 2018;2018(1):4714173.
38. Eini M.R, Darand M, Malekzadeh F, Kabiri N, Panahi A, Piniewski M. Characterization of precipitation concentration indicators and their variations in a central european region. 2023. https://doi.org/10.21203/rs.3.rs-2991363/v1.
39. Earth Engine Data Catalog: SRTM Digital Elevation Data Version 4. 2002. https://developers.google.com/earth-engine/datasets/catalog/CGIAR_SRTM90_V4. Accessed 20 July 2023.
40. Amjad M, Yilmaz MT, Yucel I, Yilmaz KK. Performance evaluation of satellite-and model-based precipitation products over varying climate and complex topography. J Hydrol. 2020;584:124707.
41. Islam T, Rico-Ramirez MA, Han D, Srivastava PK, Ishak AM. Performance evaluation of the trmm precipitation estimation using ground-based radars from the gpm validation network. J Atmos Solar Terr Phys. 2012;77:194–208.
42. Amjad M, Yilmaz MT, Yucel I, Yilmaz KK, Öztürk K. Investigating the comparative utility of ecmwf precipitation forecasts as an alternative to reanalysis data. Hydrol Sci J. 2024;69(8):1060–81.
43. Hijmans RJ, Bivand R, Forner K, Ooms J, Pebesma E, Sumner MD. Package "terra." Maintainer: Vienna, Austria; 2022.
44. Singh U, Srivastava PK, Pandey DK, Chaurasia S, Gupta DK, Chaudhary SK, Prasad R, Raghubanshi A. Scatsat-1 leaf area index product: models comparison, development, and validation over cropland. IEEE Geosci Remote Sens. 2019;17(4):563–7.
45. Baboo DSS, Devi MR. An analysis of different resampling methods in Coimbatore, district. Glob J Comput Sci technol. 2010;10(15):61–6.
46. Fadnavis S. Image interpolation techniques in digital image processing: an overview. Int J Eng Res. 2014;4(10):70–3.
47. Perich G, Turkoglu MO, Graf LV, Wegner JD, Aasen H, Walter A, Liebisch F. Pixel-based yield mapping and prediction from sentinel-2 using spectral indices and neural networks. Field Crops Res. 2023;292:108824.
48. Burger W, Burge MJ. Digital image processing: an algorithmic introduction. New York: Springer Nature; 2022.
49. Bezdek JC. Pattern recognition with fuzzy objective function algorithms. New York: Springer Nature; 2013.
50. Sardá-Espinosa A. Comparing time-series clustering algorithms in r using the dtwclust package. R package vignette. 2017;12:41.
51. Sarda-Espinosa A, Sarda MA, LazyData T. Package 'dtwclust'. Pobrane z. 2018. https://cran.r-project.org/web/packages/dtwclust/dtwclust.pdf.
52. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. 2016; 785–794.
53. Gupta HV, Kling H, Yilmaz KK, Martinez GF. Decomposition of the mean squared error and nse performance criteria: implications for improving hydrological modelling. J Hydrol. 2009;377(1–2):80–91.
54. Nash JE, Sutcliffe JV. River flow forecasting through conceptual models part i-a discussion of principles. J Hydrol. 1970;10(3):282–90.
55. Gutmann ED, Hamman JJ, Clark MP, Eidhammer T, Wood AW, Arnold JR. En-gard: a statistical downscaling framework to produce and test large ensembles of climate projections. J Hydrometeorol. 2022;23(10):1545–61.
56. Huffman G, Stocker E, Bolvin D, Nelkin E, Tan J. Gpm imerg final precipitation l3 1 month 0.1 degree× 0.1 degree v06. Goddard Earth Sciences Data and Information Services Center (GES DISC): Greenbelt, MD, USA. 2019.
57. Navarro A, García-Ortega E, Merino A, Sánchez JL, Kummerow C, Tapiador FJ. Assessment of imerg precipitation estimates over Europe. Remote Sens. 2019;11(21):2470.
58. Nascimento GJ, Althoff D, Bazame CH, Neale MUC, Duarte NS, Ruhoff LA, Gonçalves Z. Evaluating the latest imerg products in a subtropical climate: the case of Paraná State, Brazil. Remote Sens. 2021;13(5):906.
59. Risser MD, Wehner MF. The effect of geographic sampling on evaluation of extreme precipitation in high-resolution climate models. Adv Stat Climatol Meteorol Oceanogr. 2020;6(2):115–39.
60. Duchon CE, Essenberg GR. Comparative rainfall observations from pit and aboveground rain gauges with and without wind shields. Water Resour Res. 2001;37(12):3253–63.
61. Mölg T, Chiang JC, Gohm A, Cullen NJ. Temporal precipitation variability versus altitude on a tropical high mountain: observations and mesoscale atmospheric modelling. Q J R Meteorol Soc. 2009;135(643):1439–55.
62. Noor R, Arshad A, Shafeeque M, Liu J, Baig A, Ali S, Maqsood A, Pham QB, Dilawar A, Khan SN, et al. Combining aphrodite rain gauges-based precipitation with downscaled-trmm data to translate high-resolution precipitation estimates in the Indus basin. Remote Sens. 2023;15(2):318.
63. Wang Z, Kumar J, Weintraub-Leff S.R, Todd-Brown K, Mishra U, Sihi D. Upscaling soil organic carbon measurements at the continental scale using multivariate clustering analysis and machine learning. Authorea Preprints. 2023.

Publisher's Note