Ain Shams Engineering Journal 15 (2024) 102578

Contents lists available at ScienceDirect

Ain Shams Engineering Journal

journal homepage: www.sciencedirect.com

Civil Engineering

Application of multiple spatial interpolation approaches to annual rainfall data in the Wadi Cheliff basin (north Algeria)

Mohammed Achite [a,b], Paraskevas Tsangaratos [c], Gaetano Pellicone [d], Babak Mohammadi [e,*], Tommaso Caloiero [d]

[a] Faculty of Nature and Life Sciences, Laboratory of Water & Environment, University Hassiba Benboual of Chlef, P.Box 78C, Ouled Fares, Chlef 02180, Algeria
[b] National Higher School of Agronomy, ENSA, Hassan Badi 16200, El Harrach, Algiers, Algeria
[c] School of Mining and Metallurgical Engineering, Laboratory of Engineering Geology – Hydrogeology, National Technical University of Athens, Athens, Greece
[d] National Research Council of Italy, Institute for Agriculture and Forestry Systems in the Mediterranean (CNR-ISAFOM), Rende (CS), Italy
[e] Department of Physical Geography and Ecosystem Science, Lund University, Sölvegatan 12, SE 223 62 Lund, Sweden

ARTICLE INFO

Keywords:
Algeria
Deterministic techniques
Geostatistical analysis
Machine learning
Rainfall
Spatial interpolation

ABSTRACT

This study addresses a challenging problem of predicting mean annual precipitation across arid and semi-arid areas in northern Algeria, utilizing deterministic, geostatistical (GS), and machine learning (ML) models. Through the analysis of data spanning nearly five decades and encompassing 150 monitoring stations, the result of Random Forest showed the highest training performance, with R square value (of 0.9524) and the Root Mean Square Error (of 24.98). Elevation emerges as a critical factor, enhancing prediction accuracy in mountainous and complex terrains when used as an auxiliary variable. Cluster analysis further refines our understanding of station distribution and precipitation characteristics, identifying four distinct clusters, each exhibiting unique precipitation patterns and elevation zones. This study helps for a better understanding of precipitation prediction, encouraging the integration of additional variables and the exploration of climate change impacts, thereby contributing to informed environmental management and adaptation strategies across diverse climatic and terrain scenarios.

1. Introduction

Precipitation distribution exhibits substantial variability, posing a formidable challenge due to the limitations of rain gauge networks that provide sparse and discontinuous point measurements [1–3]. Addressing this challenge requires the ability to predict rainfall across unsampled locations, achieved through spatial interpolation techniques that transform point data into continuous surfaces [4,5]. Consequently, the spatial variability of rainfall emerges as a pivotal concern, particularly in the context of estimating and simulating climate change models within diverse environmental settings [6]. This challenge holds particular significance in countries characterized by arid and semi-arid climates, such as Algeria, where rainfall distribution serves as a fundamental input for various scientific applications. These applications span from watershed-scale hydrological modeling [7] to simulations informing crop growth and agricultural practices [8].

In the past two decades, significant efforts have been devoted to developing and evaluating various methodologies for spatially interpolating precipitation data, aiming to identify the most accurate and consistent approach for different geographical regions and spatial domains [9]. The emphasis on accuracy underscores the recognition that the performance of interpolation methods can vary substantially across diverse areas [10]. This concern has led to a numerous studies conducted worldwide which involved comparing and assessing different techniques for rainfall estimation. A comprehensive review by Li and Heap (2011) identified over fifty studies that compared more than seventy distinct spatial interpolation models. In the context of the Mediterranean region, where precise rainfall data is essential, di Piazza et al. (2011) conducted an extensive comparison of spatial interpolation methods. Their study, focused on creating comprehensive monthly rainfall time series for Sicily in southern Italy, examined various techniques, including Inverse Distance Weighting (IDW), Simple Linear Regression (SLR), Multiple Regression (MR), Geographically Weighted Regression (GWR), Artificial Neural Networks (ANN), Ordinary Kriging

* Corresponding author.
E-mail addresses: m.achite@univ-chlef.dz (M. Achite), tsangaratosparis@gmail.com (P. Tsangaratos), gaetano.pellicone@isafom.cnr.it (G. Pellicone), babak.mohammadi@nateko.lu.se (B. Mohammadi), tommaso.caloiero@isafom.cnr.it (T. Caloiero).

https://doi.org/10.1016/j.asej.2023.102578
Received 14 June 2023; Received in revised form 26 September 2023; Accepted 2 November 2023
Available online 25 November 2023

(OK), and Residual Ordinary Kriging (ROK). Their findings highlighted the superior performance of Residual Ordinary Kriging for both monthly and annual scales [11]. Similarly, in another region of southern Italy, Pellicone et al. (2018) conducted a detailed investigation into rainfall spatial distribution. They compared deterministic methods, such as Inverse Distance Weighting (IDW), with various geostatistical (GS) approaches, including Ordinary Kriging (OK), Kriging with External Drift (KED), Co-Kriging (COK), and Empirical Bayesian Kriging (EBK). Their study underscored the effectiveness of GS techniques over deterministic ones, with Kriging with External Drift exhibiting the lowest prediction errors [12].

Rainfall and its spatial distribution, which is considered as an effective feature in the study of climate patterns, have been also widely studied in Algeria at various scales. Keblouti et al. (2012) performed a comparative analysis among the IDW, the spline, and the OK to evaluate the most suitable method for the An-naba region, northeast Algeria, considering annual rainfall data measured in the period 1980–2008 [13]. Results evidenced that the IDW achieved the best results in the characterization of the rainfall distribution in the study area. In a study conducted by Bachir et al. (2016), which focused on the Eastern High Plateaus region of Algeria and spanned the years 1986 to 2007, a novel methodology was employed to characterize annual rainfall. This approach integrated a GS technique and MLR, taking into account pertinent geographical parameters such as altitude, longitude, and latitude, which exhibit direct correlations with rainfall [14]. Rata et al. (2020) conducted a spatial analysis of annual rainfall data covering a span from 1972 to 2012 across 58 meteorological stations situated in the Cheliff watershed, by applying the OK, the regression-kriging (RK), and the KED algorithms [15]. Cross-validation results identified the KED as the least biased interpolator, presenting few underestimates or overestimates of the data, followed by the RK and the OK.

In this context the presented study, offers a novel exploration of rainfall spatial distribution and interpolation methods, with a particular emphasis on Algeria, the Wadi Cheliff, a region characterized by arid and semi-arid climates. It addresses the challenge of accurately estimating rainfall in areas where traditional rain gauge networks provide limited and sporadic data. Notably, the study conducts a comprehensive comparison of various interpolation techniques, spanning deterministic, GS, machine learning (ML), and statistical approaches. The primary focus of the research centers on the Wadi Cheliff basin in Algeria, allowing for an in-depth examination of annual rainfall patterns over nearly five decades, from 1970 to 2018, using 150 stations. This long-term perspective offers insights into climate trends and variations, crucial for understanding the region's hydrology and water resource management. The manuscript's uniqueness stems from its detailed investigation of rainfall in Algeria, the diverse set of interpolation methods employed, and its practical implications for climate analysis and water resource management in the Wadi Cheliff basin. Additionally, while not explicitly mentioned as a novelty, the potential consideration of climate change's impact on rainfall patterns further enhances the manuscript's relevance in addressing pressing environmental challenges.

2. Study area and data used

The Wadi Cheliff, spanning approximately 750 km in length, plays a pivotal role in the socio-economic development of Algeria's primary regions. Originating in the Saharan Atlas mountains near Aflou, within the Jebel Amour region, this vital watercourse flows through a geographically diverse Wadi Cheliff Basin (WCB). The basin extends from approximately 0°7′44″ to 3°31′7″E in longitude and 33°53′13″ to 36°26′34″N in latitude, encompassing an expansive area exceeding 44,000 km2 (Fig. 1). This region exhibits a complex and rugged topography, with elevations varying from sea level to a peak of 1,969 m.

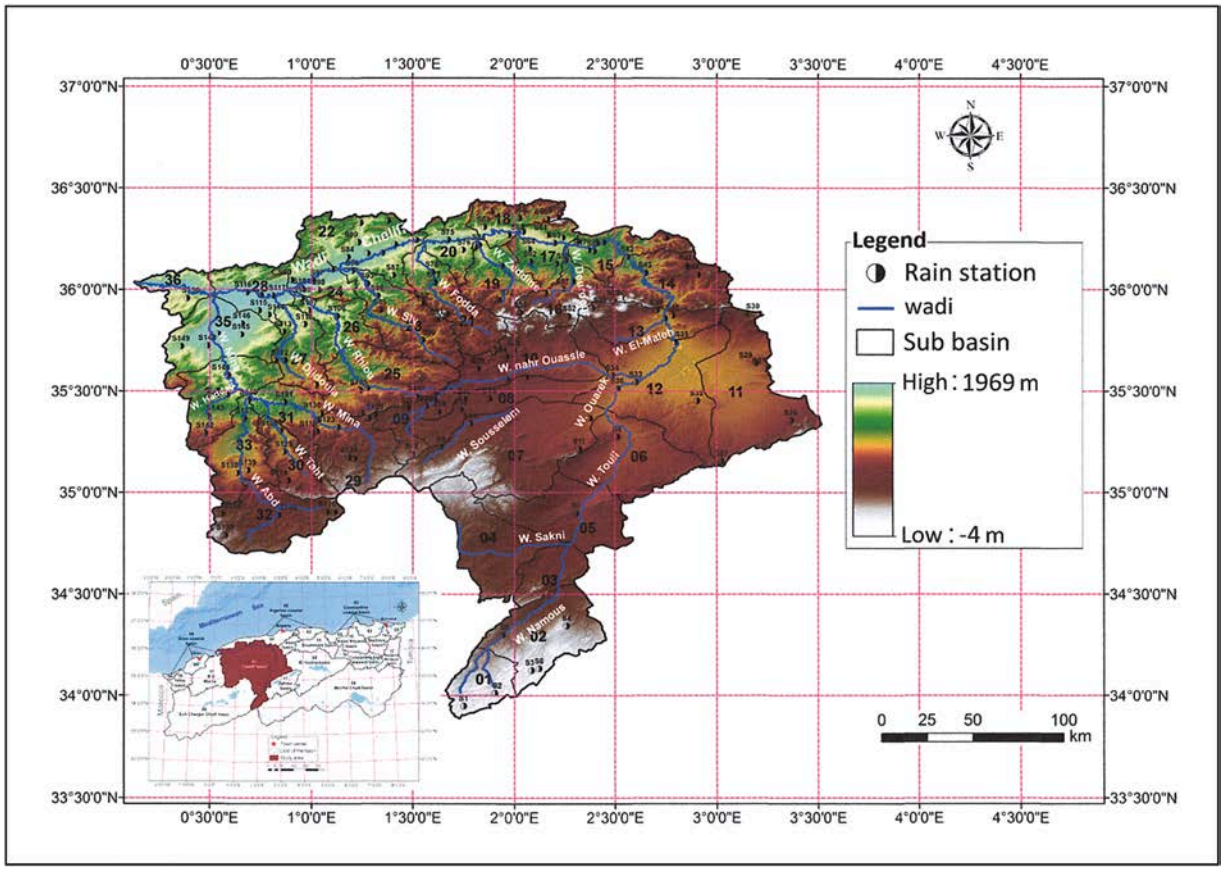

Fig. 1. Map of Wadi Cheliff basin.

The river's tributaries are symmetrically distributed from south to north, following major secondary wadis [16].

Climatically, the region belongs to the arid and semi-arid regions. A north–south gradient characterizes the mean annual temperature, with a minimum of 14.2 °C and a maximum of 18.7 °C recorded in the Tissimsilt and in the Chlef region, respectively [17]. The extreme temperatures show maximum and minimum values in July (42 °C) and January (−5 °C), respectively. The mean annual rainfall registered in the period 1970–2017 ranges between 161 mm and 662 mm, primarily concentrated during the autumn and winter months, accounting for approximately 80 % of the total precipitation.

In this study, monthly precipitation data from 150 stations managed by the National Agency of the Water Resources (ANRH) in the WCB have been considered for the period 1970 to 2017 (Fig. 1). This time interval is the longest possible, considering the availability of precipitation data in the major number of stations in the WCB. This long-term perspective offers insights into climate trends and variations, crucial for understanding the region's hydrology and water resource management.

Based on the characteristics of the research area, the intention of the study is to investigate and accurately estimate the rainfall spatial distribution of Wadi Cheliff Basin, an arid and semi-arid climate region with limited and sporadic rain gauge data. The study aims to provide valuable insights into climate patterns and fluctuations, ultimately enhancing our understanding of hydrological processes and the effective management of essential water resources in this specific region.

3. Methods

Four phases characterize the proposed methodology: (a) the pre-processing data, (b) the implementation of the GS and ML techniques and methods, (c) the comparison and evaluation of their performance, and (d) the calculation of the spatial statistical indexes to identify trends and spatial distribution. The first set of models which were used to implement the GS models was performed by applying the gstat package developed in the R environment. The ML techniques were performed by using scripts written by the programming language R packages such as caret [18]. Concerning the calculation of the spatial statistical indexes, ArcGIS 10.3.1 was the main GIS software that was used. Fig. 2 presents a flowchart summarizing the proposed methodology that is described in detail, for each phase, in the subsections as follows.

3.1. Pre-processing data

The first step of this phase, the pre-processing data, involved the choice of representative monitoring data sites. In this study, monthly precipitation data from 150 stations managed by ANRH and falling within the WCB have been considered. The next step was to separate randomly the data into training (n = 100) and test (n = 50) subsets. Finally, three independent variables (latitude, longitude, and altitude) and one dependent variable (annual rainfall) have been selected.

3.2. Implementing the geostatistical and machine learning techniques and methods

3.2.1. Inverse distance weighting (IDW)

The IDW can be considered the most applied deterministic method. The hypothesis of IDW method is based on the unknown value at an unmeasured point that can be calculated as the distance-weighted average of the known values at measured points as follows:

$$\hat{v}_i = \frac{\sum_{i=1}^{n} \frac{1}{d_i^p} v_i}{\sum_{i=1}^{n} \frac{1}{d_i^p}} \quad (1)$$

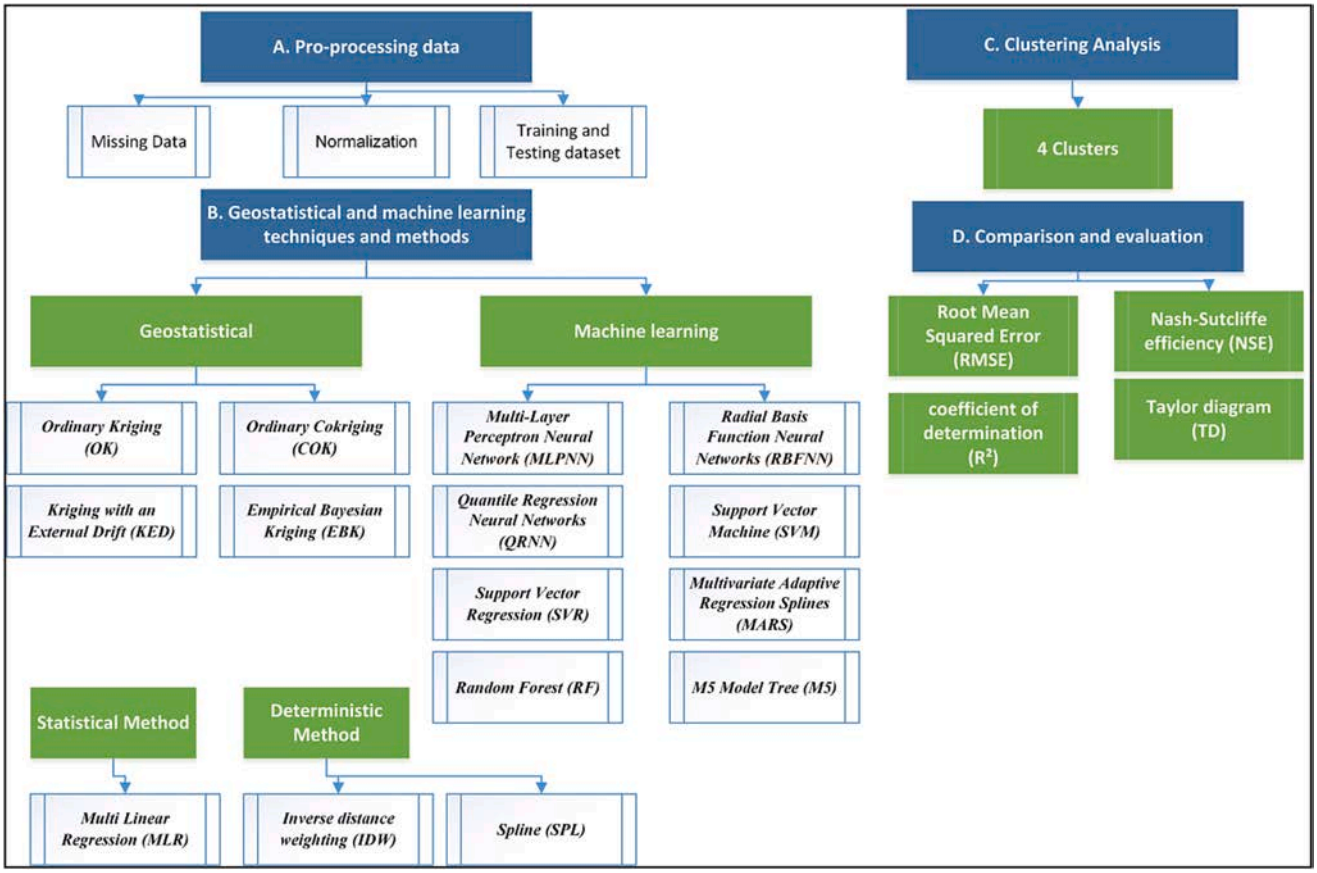

Fig. 2. Flowchart of the proposed framework in the current study.

where $\widehat{v}_i$ refers to unknown values, v_i refers to measured values, $d_i^p \ldots, d_n^p$ are distances between n data points and the point to be predicted (to the power of p).

3.2.2. Spline (SPL)

Spline interpolation is characterized by a piecewise polynomial, namely a spline as interpolant. Spline interpolation is able to recreate the value at each curve-fitting point but retains a gradient that varies uniformly [19]. A surface spline is a mathematical instrument, based on the equation of the small deflection of an infinite plate, allowing the interpolation of a function of two variables. In particular, a spline is a set of degree polynomials that are connected uniformly at certain data points. At each data point, two polynomials are connected, which results in two same-value derivatives or tangent vectors. Another requirement is that all their derivatives up to the (k − 1)st must be the same at the point. A linear spline denotes a set of segments, connecting two neighboring data points (V_k, I_k) and ($V_k + 1$, $I_k + 1$), which equation is:

$$I_k(V) = I_k + (I_k + 1 - I_k)(V - V_k)/(V_k + 1 - V_k) \tag{2}$$

with $V = [V_k, V_k + 1]$ and $k = 0,1,\ldots,(n\text{-}1)$, and by using a surface spline, the function may be differentiated in slopes finding, because it isn't required for the coordinates of the known points to be located in a rectangular array [20].

3.2.3. Ordinary kriging (OK)

Among the most common and simplest kriging methods is ordinary kriging. This first requires a cartographic analysis, and thus an experimental variogram and a variogram model.

The variogram $\gamma(h)$, which is part of the experiment, is discrete and determined via the lag h measured function by considering the variance between $N(h)$ pairs of points $[z(x_\alpha), z(x_\alpha + h)]$ at several distances:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{\alpha=1}^{N(h)} [z(x_\alpha) - z(x_\alpha + h)]^2 \tag{3}$$

Given the variogram for the random variable $Z(x)$ at a point of a region x_0, and using known data in the neighborhood $Z(x_\alpha)$ of x_0, the ordinary kriging allows us to estimate a value $Z^*_{\mathrm{OK}}(x)$ of the random variable $Z(x)$ as follows:

$$Z^*_{OK}(x_0) = \sum_{\alpha=1}^{n(x_0)} \lambda_\alpha^{OK} z(x_\alpha) \tag{4}$$

where λ_α refers to weights and $n(x_0)$ is the number of data neighboring that should be calculated.

3.2.4. Ordinary Cokriging (COK)

The COK represents an evolution of the well-known kriging algorithm allowing to estimate a value in an unsampled point by means of an auxiliary variable. Indeed, this approach allows us to exploit the covariance between two or more interconnected regionalized variables especially when sparse main features of interest and abundant related secondary data are available. For these reasons, although the COK presents the same requirements as kriging, more variography, modeling and computation time are required.

By using the COK the unknown value z_0^* is considered as a linear combination of the values of two or more regionalized variables as follows:

$$z_0^* = \sum_{i=1}^{n} \lambda_i z_i + \beta_0 t_0, \ \ldots \tag{5}$$

where λ_i is a weight (0–100 %) assigned to z_i (localized variable) at a given location, and β_j is a weight (0–100 %) assigned to t_j (secondary localized variable co-located with z_i).

3.2.5. Kriging with an External Drift (KED)

The KED is an upgrade of the kriging algorithm, its main benefit consists in the fact that it uses additional variables, correlated with the primary one, as the covariable to participate in the prediction processes. Specifically, if the primary variable presents known values in only a few points of the study site, the KED allows the calculation of the values in unsampled points using an additional variable, systematically known across the same area. Given the external deterministic variable $z(x)$ known everywhere, the model for the random function $Y(x)$ can be written as:

$$Y(x) = a + b \cdot z(x) + Y_R(x) \tag{6}$$

with $Y_R(x)$ residual stationary random field and a and b coefficients to be predicted. The estimate value $Y^*(x)$ of the random variable Y at an unknown location x_0, is a linear combination of the data:

$$Y^*(x_0) = \sum_{i=1}^{n} \lambda_i(x_i) \tag{7}$$

where $Y(x_i)$ refers to values at the n measured sites and λ_i helps to solve the following kriging system:

$$\begin{cases} \sum_{j=1}^{n} \lambda_j C(x_i - x_j) - \mu_1 - \mu_2 z(x_i) = C(x_i - x_0) \ \ for\ i = 1, \ldots, n \\ \sum_{i=1}^{n} \lambda_i = 1 \\ \sum_{j=1}^{n} \lambda_i z(x_i) = z(x_0) \end{cases} \tag{8}$$

where C id covariance function and μ_1 and μ_2 are Lagrange parameters.

3.2.6. Empirical Bayesian kriging (EBK)

The EBK hybrids Bayes' theorem and kriging technique taking the parameters as random variables [21]. With slightly non-stationary data, the EBK is increasingly more flexible than other kriging methods and needs fewer presumptions about the modeled semivariogram [22]. In effect, by using the EBK the most intricate facets of constructing a sound kriging model by sub-setting the study area, coupled with several simulations implemented by iteratively estimating., a lot of semivariogram models to obtain the best fit, can be automated [23]. As with other kriging methods, the EBK uses an intrinsic random function, though differing from the others by using the error which introduced in the semivariogram prediction, thus considering for the observations' uncertainty.

The EBK requires the following steps [24]:

1. Estimation of a semivariogram model from known data;
2. Simulating new values for every data point of the subset by using the semivariogram which was estimated in the first step;
3. Calculation of a new semivariogram model with the data of step 2.

In the step 3, the weight of the semivariogram is computed by applying the Bayes' rule: the latter measures how probable an estimated semivariogram simulates observed data. Steps 2 and 3 must be repeated 150 times (as the number of the known values) to evaluate estimations and their standard errors at the places without samples, using the weights calculated in step 3. The final result of this procedure is a variety of semivariograms providing an estimation of the real semivariogram for the subset. Finally, the true semivariogram model for that subset is obtained from the weighted sum of all simulated semivariograms. This method is used to calculate the likelihood of a particular simulation in terms of the density of probabilistic distributions for various parameter values.

3.2.7. Multi-Layer Perceptron Neural Network (MLPNN)

The Multi-Layer Perceptron Neural Network (MLPNN) is a powerful ML model commonly used for both regression and classification tasks. It consists of multiple layers, including an input layer, one or more hidden layers, and an output layer with each layer contains interconnected nodes, often referred to as neurons [25]. Concerning the architecture of the MLPNN, the input layer consists of nodes equal to the number of independent variables used in the model. These nodes receive the initial input data. It typically includes one or more hidden layers, each containing a variable number of neurons with the neurons interconnected with both the input and output layers. The output layer produces the final results. The number of nodes in this layer depends on the nature of the problem—regression tasks may have a single output node, while classification tasks can have multiple nodes, each representing a class [26].

Each connection between nodes in different layers is associated with a weight coefficient. These weights determine the strength of the connection and are adjusted during the training process to optimize the model's performance [27].

Concerning the training phase, a typical MLPNN model uses a back-propagation learning algorithm involving three stages: the feed-forward, the calculation and back-propagation of the associated error and the adjustment of the weights stage. The hidden and output layer nodes process their inputs with a nonlinear transfer function, usually a sigmoid function, to calculate a result [28,29]:

$$f(net) = \frac{1}{1 + e^{(-net)}} \tag{9}$$

$$net_j^l(t) = \sum_{i=o}^{p} \left(y_i^{l-1}(t) w_{ji}^l(t) \right) \tag{10}$$

The error is calculated by comparing the predicted values to the actual observed values. This error is then propagated backward through the network. During this back-propagation step, the weights of connections are adjusted to minimize the error and improve the prediction accuracy. The training phase involves adjusting the weight coefficients for each connection with key parameters in this process including the learning rate, learning decay rate, and momentum [29]. The learning rate controls the magnitude of weight updates, the learning decay rate decreases the learning rate over iterations, and momentum helps maintain the direction of weight updates. The training phase continues until a predefined stopping criterion is met. Typically, it ends when the model achieves minimal error on the training data, indicating that it has learned the underlying patterns in the data. Finally, after successful training, the MLPNN can be used for making predictions on new, unseen data, providing valuable insights for various applications. A grid search approach has been applied to tune the hyper parameters, the number of neurons in each layer and learning rate.

3.2.8. Radial basis function Neural Networks (RBFNN)

Similar to the MLPNN models, RBFNN are composed by an input, a hidden and an output layer. The difference between the two models is that RBFNN uses radial basis functions (RBF) as activation functions. The most common RBF is the Gaussian function [30]:

$$RBF_i(X) = \exp\left(-\frac{\|X - c_i\|^2}{2\sigma^2} \right) \quad i = 1, 2, ..., m \tag{11}$$

where X is an n-dimensional vector, $RBF_i(X)$ is the activation function, m the number of hidden nodes, c_i the center of the i^{th} basic function and $\|X - c_i\|$ is the Euclidian norm of $X - c_i$ and σ is the width of the hidden units.

3.2.9. Quantile regression Neural Networks (QRNN)

The QRNN has been proposed by Taylor (2000) and refers to as a flexible non-linear type of quantile regression to estimate the conditional quantiles of a predictive parameter which rely on covariates in some form of regression equation [31]. The QRNN has a similar topology as the MLPNN model [31–33]. In the case of a set of input variables $x_i(t)$ and an output $y(t)$, the result of the implementation of QRNN, is as follows (Cannon, 2011). The output $g_i(t)$ from the j^{th} hidden-layer node is equal to:

$$g_i(t) = \tanh\left(\sum_{i=1}^{I} x_i(t) w_{ij}^{(h)} + b_j^{(h)} \right) \tag{12}$$

with $w_{ij}^{(h)}$ and $b_j^{(h)}$ hidden-layer weights and the hidden-layer bias, respectively.

An estimate of the conditional τ-quantile ($\hat{y}_\tau(t)$) is then provided by following formula:

$$\hat{y}_\tau(t) = f\left(\sum_{i=1}^{J} g(t) w_j^{(0)} + b^{(0)} \right) \tag{13}$$

where $w_j^{(0)}$, $b^{(0)}$, and $f()$ are the output-layer weights, bias, and transfer function, respectively.

3.2.10. Support Vector Machine (SVM)

SVM is also known as non-parametric kernel-based techniques that was suggested by Vapnik et al. [34], and usually SVM applies to handle linear and non-linear classification and regression problems. SVM applies optimal linear hyperplane separating data models and implements kernel-based functions to transform the original non-linear data into a linearly separable format in high-dimensional feature space [35]. The Radial Basis Function was applied while the regularization parameter C was tuned following a grid search approach.

3.2.11. Support Vector regression (SVR)

SVR models follow, in a similar manner to SVM models, a training and prediction procedure for classification problems, but the output of their implementation is a real number [36]. In main objective of SVR models, is to detect a regression function $f(x)$ able to predict a value with a maximum of ε (*epsilon*) deviation from the actual values and, at the same time, errors lower than ε.

$$F(x) = w \cdot x - b \tag{14}$$

$$subject\ to\ y_i - w \cdot x_i - b \leqslant \varepsilon \tag{15}$$

where x_i denotes the i^h sample and y_i represents the value of the i^{th} sample. In this study, the RBF Gaussian kernel has been implemented to transform the linear regression formulation to a non-linear SVR [37].

$$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2 \right) \tag{16}$$

The main objective of the SVR is to minimize the error, between the predictive and the observed values [38]. This error is influenced by the kernel width γ (*gamma*), by ε and by the regularization parameter C (*cost*). And in this case, the Radial Basis Function was applied while the regularization parameter C, ε and γ were tuned following a grid search approach.

3.2.12. Multivariate Adaptive regression Splines (MARS)

Multivariate Adaptive Regression Splines (MARS) is a flexible algorithm used for modeling complex non-linear relationships between variables. It combines linear regression with piecewise functions, making it particularly effective when dealing with data that exhibits non-linearity and interactions between variables [39]. It operates through a two-stage process: the forward and backward stages.

In the forward stage of the MARS, the algorithm starts with a simple linear regression model. It searches for breakpoints or "knots" in the

independent variables (X) to create piecewise linear segments. These breakpoints are identified based on the data's characteristics, and they represent critical values where the relationship between the dependent variable (Y) and X changes significantly. For each identified breakpoint, MARS adds a basis function, which is essentially a piecewise linear segment.

Consider a simple model of Y ~ X. During the forward stage, MARS identifies the single value among the range of X values where two dissimilar linear functions between Y and X produce the lowest error. MARS constructs models that have the form:

$$f(x) = \sum_{i=1}^{k} c_i B_i(x) \tag{17}$$

with c_i constant coefficient and $B_i(x)$ weighted sum of basis functions. In particular, $B_i(x)$ can be a constant, a hinge function (max (0, x-constant) or max (0, constant-x)) or a product of two or more hinge functions [40].

After adding basis functions, MARS may end up with a relatively complex model. The backward stage, known as pruning, aims to simplify the model while preserving its predictive accuracy. It does this by iteratively removing basis functions that do not contribute significantly to the model's performance. The Generalized Cross-Validation (GCV) score is often used as a criterion to guide this process. GCV measures the model's goodness of fit while penalizing model complexity. Equation 18 and 19 provides the formula to calculate the GCV score.

$$GCV(M) = \frac{\frac{1}{N}\sum_{i=1}^{n}[y_i - f(x_i)]^2}{\left[1 - \frac{C(M)}{N}\right]^2} \tag{18}$$

$$\mathrm{C(M)} = (\mathrm{trace(B\ (B^T B) - 1\ B^T)} + 1) + \mathrm{dM} \tag{19}$$

where y_i is the i^{-th} observed predictive variable, N the number of observations, M the number of basic functions and C(M) represents the penalty measure for complexity, d is defined as the smoothing parameter and B the matrix of basic functions with dimension M × N.

As basic functions are removed, the model becomes simpler but still retains its ability to capture the underlying non-linear relationships in the data.

3.2.13. Random Forest (RF)

In general, RF models are characterized as ensemble models used in classification and regression problems. They produce a boosted outcome based on the combined results of a number of trained weak decision tree learners with the ability to generalize and minimize the risk of over fitting, without having to undergo any pruning process [41]. The training process involves the creation of a number of different bootstrap samples derived from the original training dataset, with a subset of samples being left out of the training process in order to form the validation dataset which allows the estimation of an unbiased test error, known as out-of-bag-error, that expresses the predictive ability of the RF model [42,43]. RF combines two different processes during the training phase: a bagging algorithm, applied to a set of tree classifiers (base models), and a random selection of the prediction variables. The main difference from a simple bagging method is based on the procedure of constructing decision trees, which are trained and evaluated on only a subset of randomly chosen features. In order to effectively implement the Random Forest (RF) model, two crucial parameters require adjustment: the number of trees employed by the model and the number of predictor variables randomly selected from the complete set of variables [43,44]. In this study, we fine-tuned both of these critical parameters using a grid search approach. This method systematically explored a range of values for each parameter, assessing their impact on model performance, and ultimately selecting the combination that yielded the best predictive results. The grid search approach enabled us to optimize the RF model's hyperparameters, ensuring its robustness and accuracy in predicting mean annual precipitation [45].

3.2.14. M5 model tree (M5)

The algorithm behind M5 tree model, proposed by Quinlan (1992), uses binary decision trees where cause and effect variables are mapped from linear regression functions in the terminal node [46,47]. M5 model tree is built on the assumption that the dependency is not constant between input and output in the whole range of data, but can be approximated as such on smaller subset of data [48]. In case of continuous parameters, subsets of data are identified applying either the average value, or with a linear regression function of the predictive parameter:

$$y = a_0 + a_1\, ax_1 + a_2\, ax_2 \tag{20}$$

The splitting criterion is based on the standard deviation of the values in the subset which is used as a metric expressing the error:

$$SDR = sd(T) - \sum \frac{|T_i|}{|T|} xsd(T_i) \tag{21}$$

in which T_i corresponds to the subsets that are products of the splitting process (T_1, T_2, T_3). The splitting process ends when the standard deviation is less by a small fraction than the standard deviation of the original dataset or in the case of the remaining few data.

3.2.15. Multi linear regression (MLR)

Given a set of independent variables, the MLR is the statistical method used to predict the values of a dependent variable with the following formula:

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + , ..., + \beta_k X_n \tag{22}$$

where Z refers to the dependent variable, $\beta_0, \beta_1, \beta_2, ..., \beta_k$ are the regression coefficients and $X_1, X_2, ..., X_n$ are the independent parameters whereas n corresponds to the number of independent variables and k to the number of regression coefficients. The degree of influence of each predictor variable on the response variable is expressed by the regression coefficients.

3.3. Comparison and evaluation of the model s ' performance

According to Legates and McCabe (1999) the most necessary performance metrics during a model assessment should include at least one goodness-of fit measure, and one absolute error measure [49]. In this study, in order to evaluate and compare the performance of the fifteen models, the RMSE, the NSE and the R^2 have been evaluated. Moreover, a visual interpretation was also made by using the Taylor diagram (TD).

3.3.1. Root mean squared error (RMSE)

The RMSE is a statistical metric that allows the detection of estimate errors as follows [50]:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(p_{io} - p_{ip})^2}{n}} \tag{23}$$

3.3.2. Nash-Sutcliffe efficiency (NSE)

Given a plot of observed versus predictive values, the NSE [49] is a measure ranging from −1 (low accurate prediction) to 1 (perfect predictive accuracy).

$$NSE = 1 - \frac{\sum_{i=1}^{n}(p_{io} - p_{ip})^2}{\sum_{i=1}^{n}(p_{io} - \overline{p_{ip}})^2} \tag{24}$$

3.3.3. Coefficient of determination

As evidenced by Meng et al. (2013) and Borges et al. (2016) the R^2 evaluates the linear dependent similarity between a measured value and an estimated one with the following formula [51,52]:

$$R^2 = \left(\frac{\sum_{i=1}^{n} (p_{io} - \overline{p_o})(p_{ip} - \overline{p_p})}{\sqrt{\sum_{i=1}^{n} (p_{io} - \overline{p_o})^2} \sqrt{\sum_{i=1}^{n} (p_{ip} - \overline{p_p})^2} } \right)^2 \quad (25)$$

3.3.4. Taylor diagram (TD)

TDs are mathematical diagrams constructed in such a way that could graphically indicate which among a number of models is more accurate or may simulate the environment in a more realistic manner. Taylor diagrams was first introduced by Taylor (2001) in order to enable a comparative assessment of different models [53]. The main objective of the TD is to provide a graphical product so as to quantify the degree of correspondence between predictive and actual data by means of three statistics: the Pearson correlation coefficient, the RMSE and the standard deviation.

3.4. Cluster analysis

A Cluster Analysis was implemented during the study based on K-means unsupervised clustering algorithm. The aim of K-means is to partition a set of samples into k clusters in which each samples belongs to the cluster with the nearest mean sample that serves as the most representative sample of the cluster. In our case, the algorithm will partition the rainfall stations into clusters that shear common characteristics. The following parameters were taken into account: observed precipitation value, maximum, minimum, SD and CV values. The optimal number (k) was set by implementing the elbow rule [54].

4. Results and discussion

As several researchers have concluded, there is no single universal method for precipitation assessment that can be applied with a high predictive performance in any case [55]. The best predictor model for a specific research area can only be found through a comparative assessment of different methods and techniques. In the present study, this was done by comparing fifteen models, which were grouped as deterministic, GS and ML models. The outcomes of the study revealed the superiority of the ML models against the GS and deterministic models, but also the influence that certain spatial characteristics and relations among geo-environmental variables had on the predictive power of the different models. Our comparative assessment of fifteen different models, categorized as deterministic, GS, and ML models, underscored the complexity of precipitation assessment. It confirmed that the choice of the most suitable model is highly context-dependent and relies on the unique characteristics of the research area. All ML models used longitude, latitude and elevation characteristics as variables, having as a target the precipitation data. Elevation emerged as a critical auxiliary variable that significantly improved precipitation estimation, particularly in regions characterized by mountainous and complex terrain. According to Adhikary et al. (2017), the use of elevation as an auxiliary variable can improve the estimation of rainfall especially in catchments characterized by mountainous and/or complex terrain [55]. In our study, among the GS models, COK and KED gave better results than other GS and deterministic models, probably due to the use of elevation as an auxiliary variable. Our study went one step further since it analyzed the gauge stations and their spatial distribution, with reference to the performance of all models. Specifically, taking into account the criterion of minimum error, a significant number of stations in which none of the used models gave satisfactory results were identified. All stations appear to share some common characteristics, which would be reported later.

In our case, although we did not produce a mean annual precipitation map, showing the spatial distribution of the predictive values, several spatial characteristics were observed. The study area has a high spatial variability concerning rainfall. Fig. 3a illustrates the mean annual precipitation distribution of the observed values concerning the 150 gauge stations of the research area, along with descriptive statistical measures distribution analysis (Fig. 3b). Higher values (644 mm) are located in the northern parts of the study area which are characterized by lower altitudes, whereas lower values (154 mm) are located at south and east parts of the study area. Most values are within the range of 294–364 mm with the distribution of the data characterized as normal. Concerning the spatial distribution of the 150 gauged stations in reference to the elevation, it was found that twenty-nine (29) stations are within elevation less than 200 m, forty-two (42) within 200 and 600 m, fifty-seven (57) station within 600 and 1,000 m and twenty-two (22) stations above 1,000 m. Spatial distribution analysis revealed substantial variability in mean annual precipitation across the study area. Higher precipitation values were concentrated in the northern regions with lower altitudes, while lower values prevailed in the southern and eastern parts. This spatial pattern suggests a strong geographical influence on precipitation patterns, with elevation acting as a key driver.

Concerning the maximum observed values and based on the entire data (1970–2017) the highest values were found in the north and mainly in the central parts of the research area. Most values appear within the range of 400–500 mm and 600–700 mm, whereas the distribution could be characterized as multi-modal distribution (Fig. 4a, b). Concerning the minimum values they are mainly observed at the south parts of the area, with most values ranging between 124 and 224 mm, exhibiting a right-skewed distribution (Fig. 4c, d). Similar pattern could be identified for the Standard Deviation (SD) parameter concerning the spatial

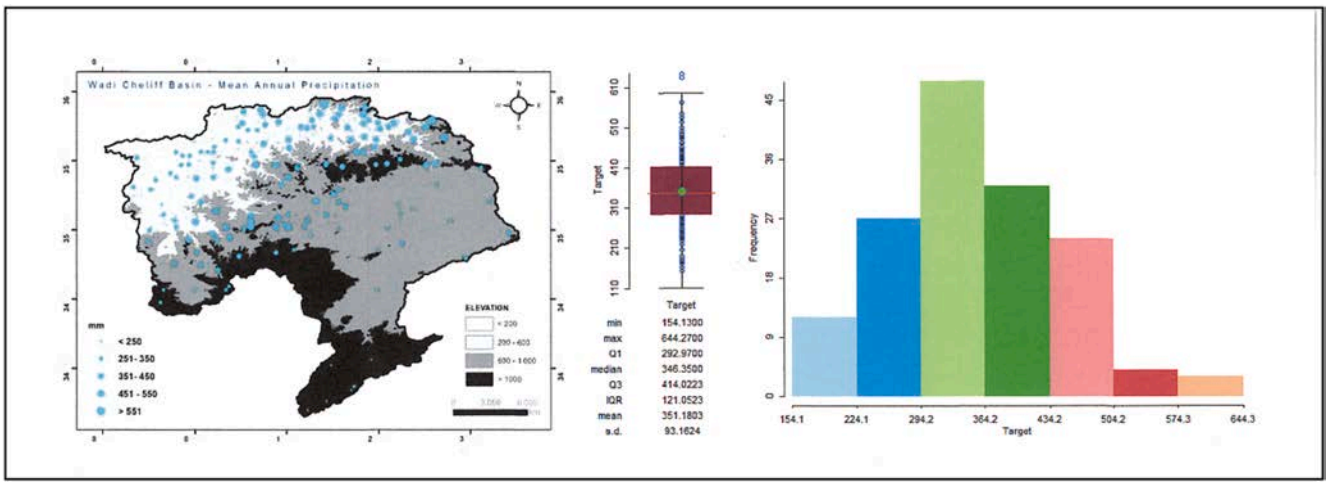

Fig. 3. Mean annual precipitation information.

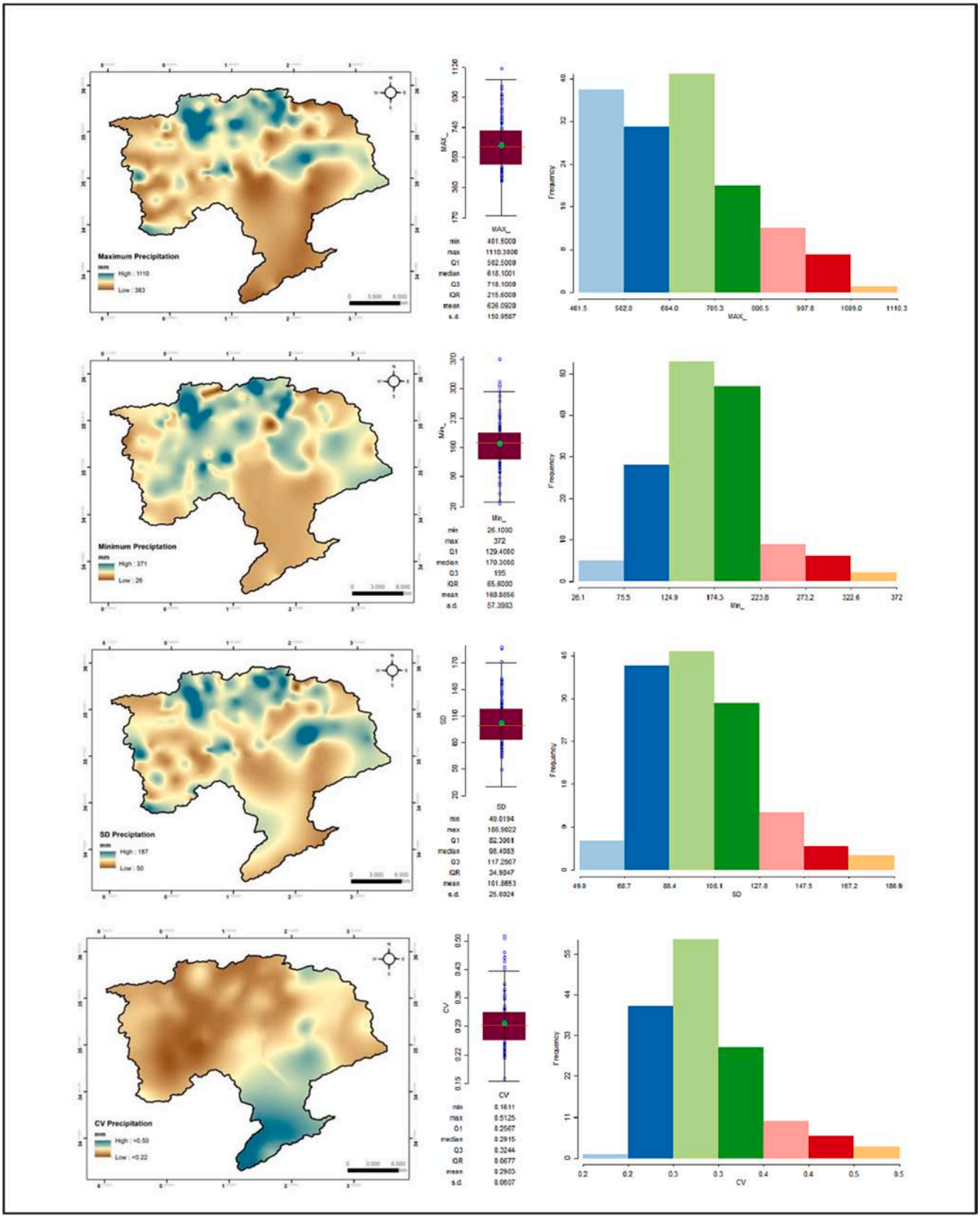

Fig. 4. Box and Whisker Plots for mean annual precipitation, maximum, minimum, SD and CV values.

distribution (Fig. 4e, f), whereas coefficient of variation (CV) has a different spatial distribution from the above (Fig. 4j, h). The highest values are observed in the southern areas, while lowest values were spread across the central and northern areas. Our outcomes are in accordance with Achite et al. (2021) who reports that (for the same research area) the CV ranges between 16.0 % (in the northern areas of the region) and 56.5 % (in the mountainous areas of the south side of the region). This variability is directly related to the geomorphological alteration that characterizes the area in which plains, valleys, plateaus and isolated or chain mountains can be observed. Specifically, plains are often associated with more consistent and predictable precipitation patterns due to their relatively uniform topography. In contrast, valleys can exhibit microclimates characterized by variations in temperature and precipitation. The convergence of winds and moisture-laden air masses in valleys can lead to localized rainfall patterns.

After performing a clustering analysis which took into account the following parameters, observed value, maximum, minimum, SD and CV values, four clusters were identified: A, B, C and D with the following characteristics (Figs. 5 and 6). Cluster A which contains 29 stations, and is characterized by mean annual precipitation of about 253 mm and values of precipitation ranging between 154 and 344 mm. It covers approximately 35.4 % of the entire area, whereas most of the gauge stations within cluster A, specifically twelve (12) are located in the elevation zone which ranges between 600 and 1,000 m. Eight (8) stations are characterized by an elevation above 1,000 m, seven (7) with an elevation less than 200 m and two (2) range between 200 and 600 m. Cluster B contains 24 stations and is characterized by mean annual precipitation of 279 mm and values ranging within 180–347 mm. It covers 21.9 % of the entire area, with eleven (11) stations having elevation less than 200 m, ten (10) between 600 and 1,000 m and three (3) between 200 and 600 m. Cluster C contains 55 and is characterized by a mean annual precipitation of 408 mm and values ranging between 288 and 644 mm. It covers 28.2 % of the entire area, with twenty-five (25) stations having elevation which ranges between 600 and 1,000 m, twenty (20) between 200 and 600 m and ten (10) above 1,000 m. Finally, cluster D contains 42 stations characterized by mean annual precipitation of about 384 mm and values ranging from 178 to 634 mm. It covers the rest 14.5 % of the entire area and seventeen (17) stations are located within the elevation zone of 200–600 m, eleven (11) are characterized by elevation less than 200 m, ten (10) with elevation between 600 and 1,000 m and four (4) above 1,000 m. From the visual inspection of the map that shows the spatial distribution of the four clusters, cluster C seems to be stretched through the entire area in a more uniform manner, than the other three clusters that can be observed in several areas in varying extent.

Based on the four statistical metrics and the training and test subsets (Table 1 and Fig. 7), the RF model showed the highest training performance, with reference to the r square value (0.9524) followed by the MLNN (0.9203), whereas the lowest values were achieved by the IDW model (0.5291) followed by the SVM model (0.5803). The RF model also showed the highest predictive performance, with reference to the r square value (0.8408), whereas the SVR showed the second highest value (0.8132). The lowest values were again achieved by the IDW model (0.4308), followed by the OK model (0.6038). Concerning the RMSE values, RF had the lowest value (24.98), followed by MLPNN (27.74), whereas IDW had the highest (71.01) followed by the SVM (63.58) based on the training subset. A different pattern has been observed concerning the test subset, with the exception of the RF model which still has the best value (33.94). The SVR model had the second lowest value (35.09), whereas the worst predictive performance was observed by the IDW model (63.98) followed by the OK model (51.70). Similarly, the RF model had the best learning and predictive performance concerning the NSE and MAE metrics (0.9352 and 19.62, 0.8244, and 26.51, respectively). Also, the learning performance of the MLPNN model was slightly worse than the RF model (0.9200 and 22.26). The SVR model also had the second predictive performance (0.8122 and

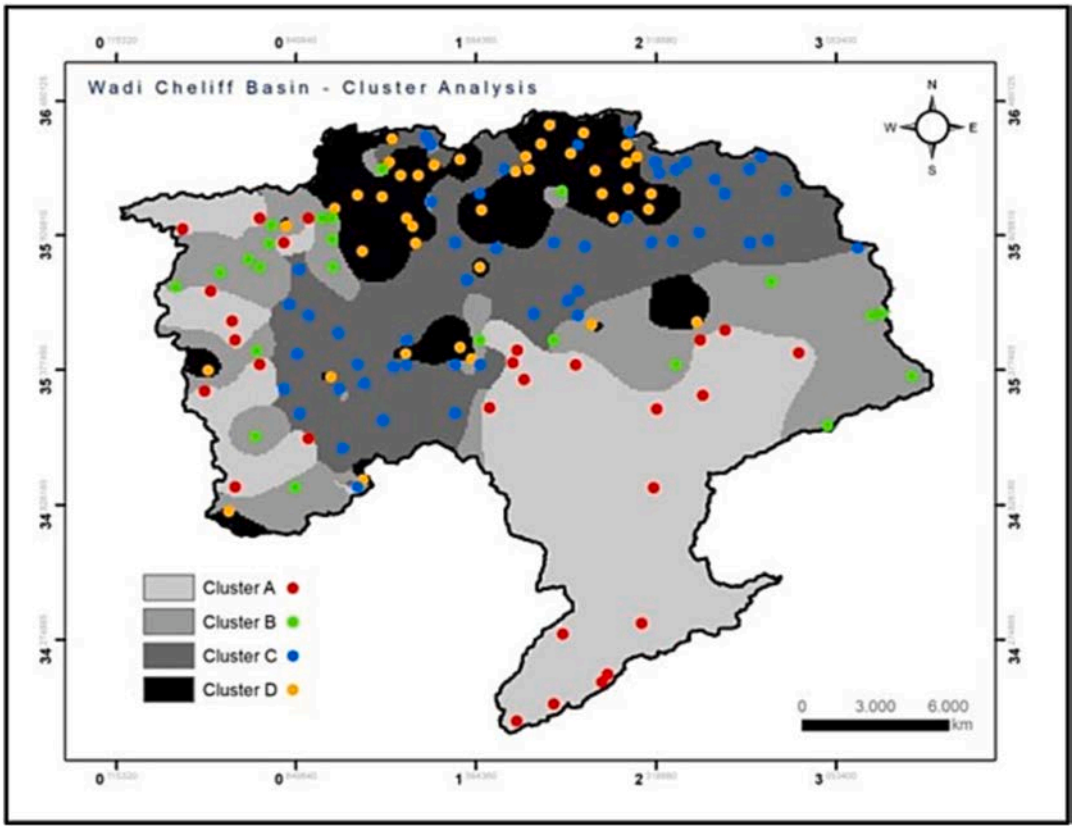

Fig. 5. Spatial distributions of clusters.

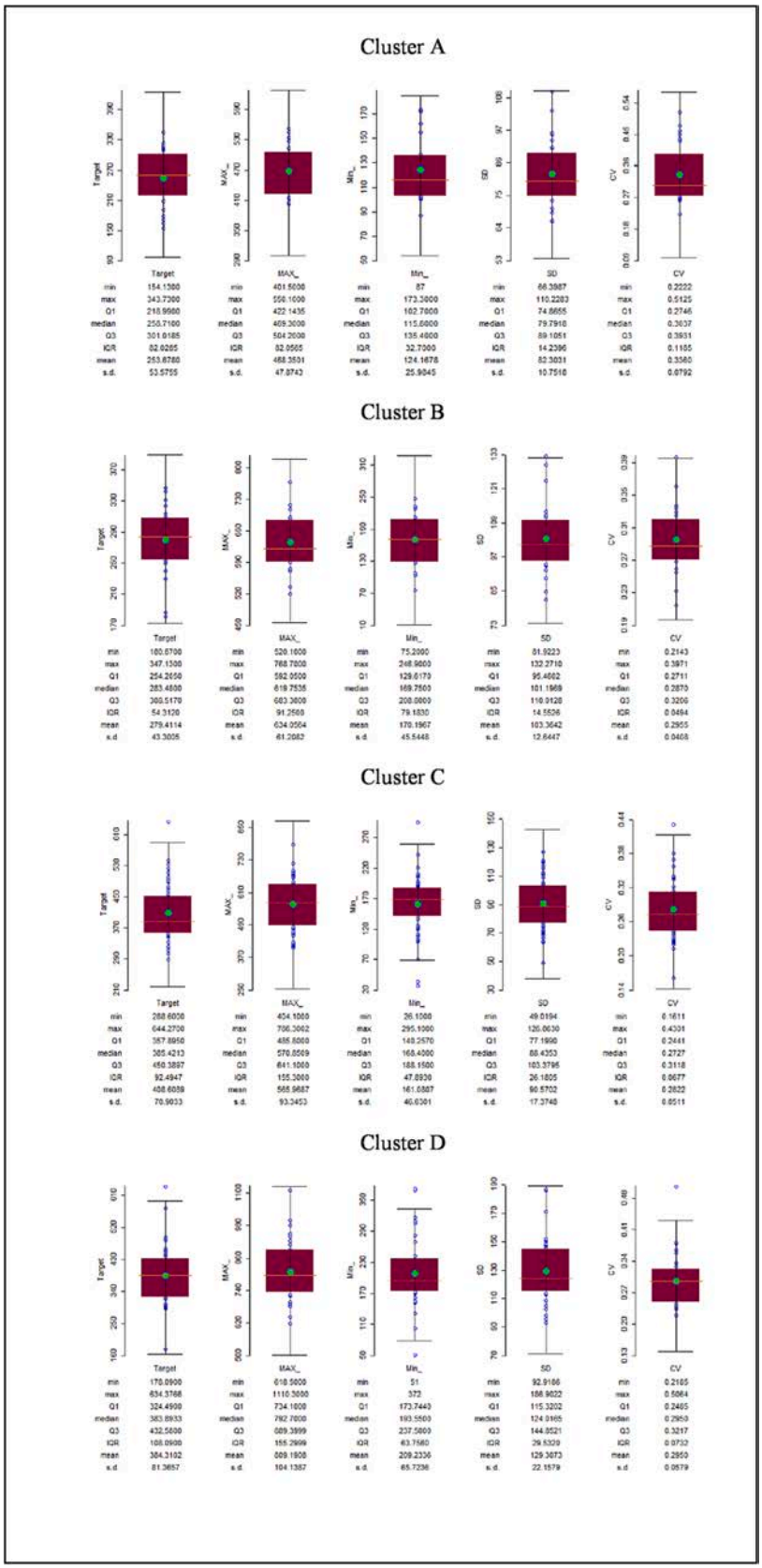

Fig. 6. Box and Whisker Plots for clusters A, B, C, and D.

Table 1
Statistical metrics of each used method for training and testing sets.

	Training Subset				Testing Subset			
	R^2	RMSE	NSE	MAE	R^2	RMSE	NSE	MAE
IDW	0.5291	71.01	0.4764	55.64	0.4308	63.98	0.3952	44.92
Spline	0.6268	60.23	0.6233	45.96	0.6048	51.61	0.5938	37.18
OK	0.6565	57.65	0.6548	44.67	0.6038	51.70	0.5925	38.10
COK	0.7147	52.44	0.7144	40.19	0.6255	50.71	0.6078	38.45
KED	0.6944	54.25	0.6943	42.76	0.7302	43.48	0.7117	31.84
EBK	0.6680	56.65	0.6670	43.58	0.6244	49.97	0.6192	35.16
MLPNN	0.9203	27.74	0.9200	22.26	0.7135	49.02	0.6334	37.38
RBFNN	0.7228	51.78	0.7215	41.05	0.7957	36.62	0.7954	29.23
QRNN	0.5874	63.40	0.5824	47.73	0.6282	50.60	0.6096	39.60
SVM	0.5803	63.58	0.5801	48.30	0.6287	50.38	0.613	38.89
SVR	0.7904	45.61	0.7838	34.35	0.8132	35.09	0.8122	27.41
MARS	0.6924	54.42	0.6924	41.79	0.6519	48.53	0.6409	32.94
RF	0.9524	24.98	0.9352	19.62	0.8408	33.94	0.8244	26.51
M5	0.7983	44.19	0.7972	34.71	0.8082	36.88	0.7925	28.41
MLR	0.5873	63.04	0.5872	49.09	0.6278	50.56	0.6101	40.52

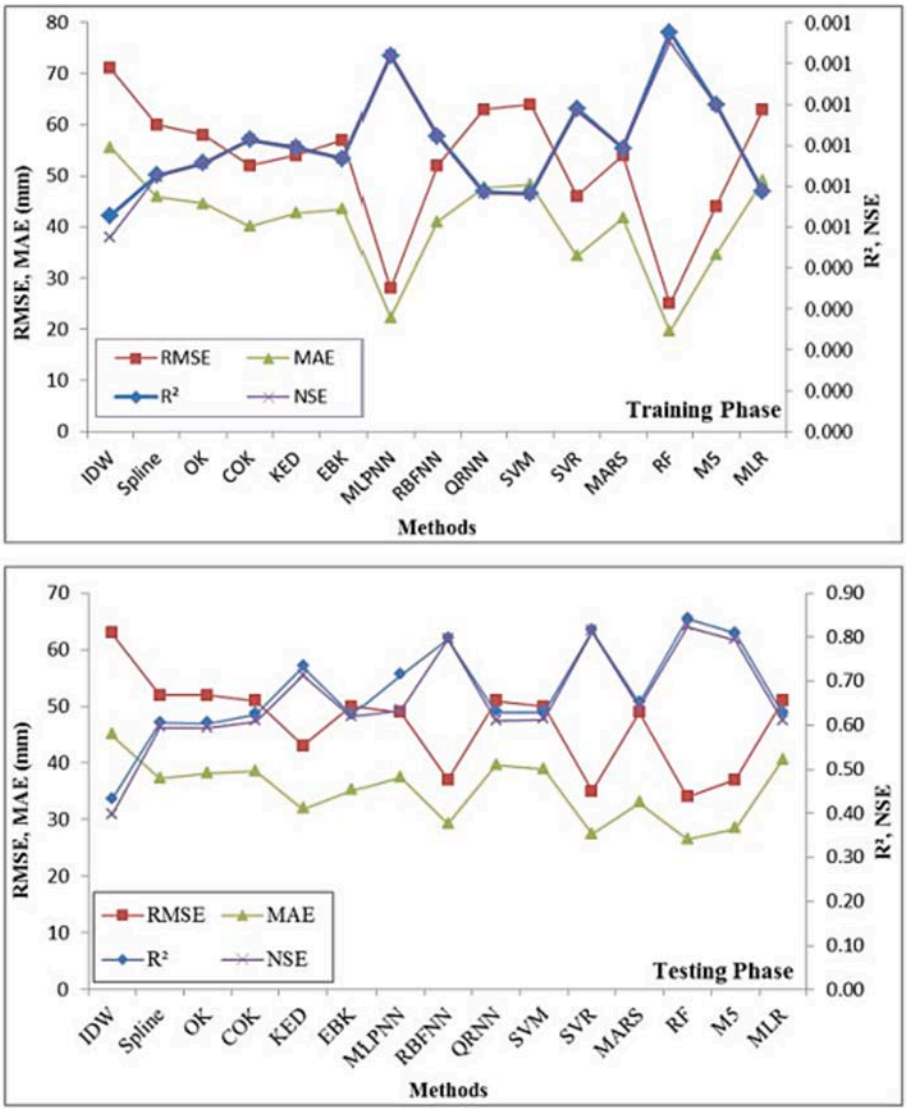

Fig. 7. Diagram of the four statistical metrics.

27.41), whereas in every case, the IDW model had the worst performance (0.4764 and 55.64, 0.3952 and 44.92, respectively).

From the Taylor diagrams (Fig. 8) one can observe the accuracy achieved by the RF model compared to the rest of the models, in both training and test subsets, whereas the IDW model had the worst performance in both cases.

Fig. 9a illustrates for each training gauge station the number of models that predicted the mean annual precipitation with an absolute

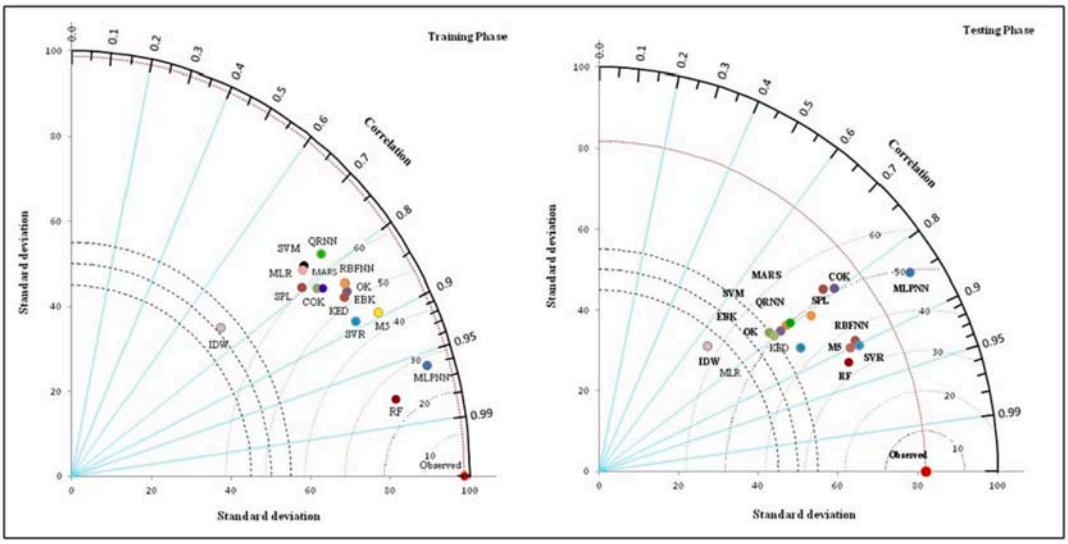

Fig. 8. Taylor Diagrams for the training and test subsets.

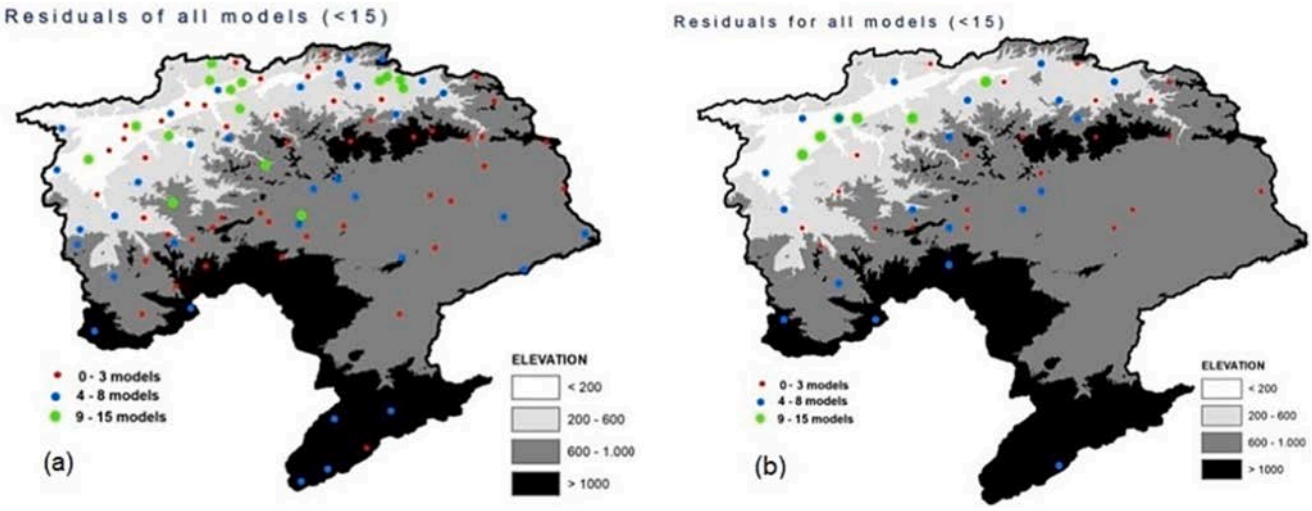

Fig. 9. Training subset/elevation – number of models (left) and testing subset/elevation – number of models (right).

difference less than 15 mm than the observed values based on the training subset. Instead Fig. 9b illustrates the same distribution based on the test subset. Most of the models within the specified residual criteria (residuals < 15 mm) performed better at the northern areas which are characterized by an elevation of less than 600 m and higher precipitation values. The same pattern has been observed when taking into account the test subset as can be seen in Fig. 9b.

More specifically, seventy-eight (78) stations had at least one ML model that meets the residual criteria, concerning the train subset and forty-one (41) stations concerning the test subset. In total, almost 79 % of the stations had at least one ML model that meets the residual criteria. Forty-nine (49) stations had at least one GS model concerning the train subset and thirty-five (35) stations concerning the test subset. In total almost 56 % of the stations had at least one GS model that meets the residual criteria. Furthermore, only a small percentage of the total number of stations (12 %) had more GS models which meet the residual criteria than ML models. It is significant to report that for twenty-five (25) stations none of the applied models gave results that captured the residual criteria. Three (3) stations belong to cluster A which corresponds to around 10.3 % of the number of stations found cluster A, six (6) to cluster B (25.0 % of the total number of stations in cluster B), ten (10) to cluster C (18.2 % of the total number of stations in cluster C) and six (6) to cluster D (14.3 % of the total number of stations in cluster D). For the same group of stations, fifteen (15) were found in the elevation zone 600–1,000 m (26.3 % of the total number of stations with the specific zone), five (5) in the elevation zone 200–600 m (11.9 % of the total number of stations with the specific zone), four (4) above 1,000 m (18.2 % of the total number of stations with the specific zone) and one (1) in the elevation zone less than 200 m (less than 1 % of the total number of stations with the specific zone). Taking into account the number of stations found in each class, identified through the cluster analysis and elevation classification, elevation zone C (600–1,000 m) and cluster B, where the classes that had the highest percentage in stations that none of the applied models gave results that captured the residual criteria. Cluster B had the second highest mean values for SD, CV, minimum and maximum and the third highest mean value for the actual mean annual precipitation. Fig. 10 illustrates the areas that meet the above criteria. Obvious these areas are characterized by the lower presence of gauge stations, thus producing less accurate results.

According to the results of our study, the best learning and predictive performance was achieved by the RF model. This could be contributed to the fact that RF builds an ensemble of trees based on the principle that a

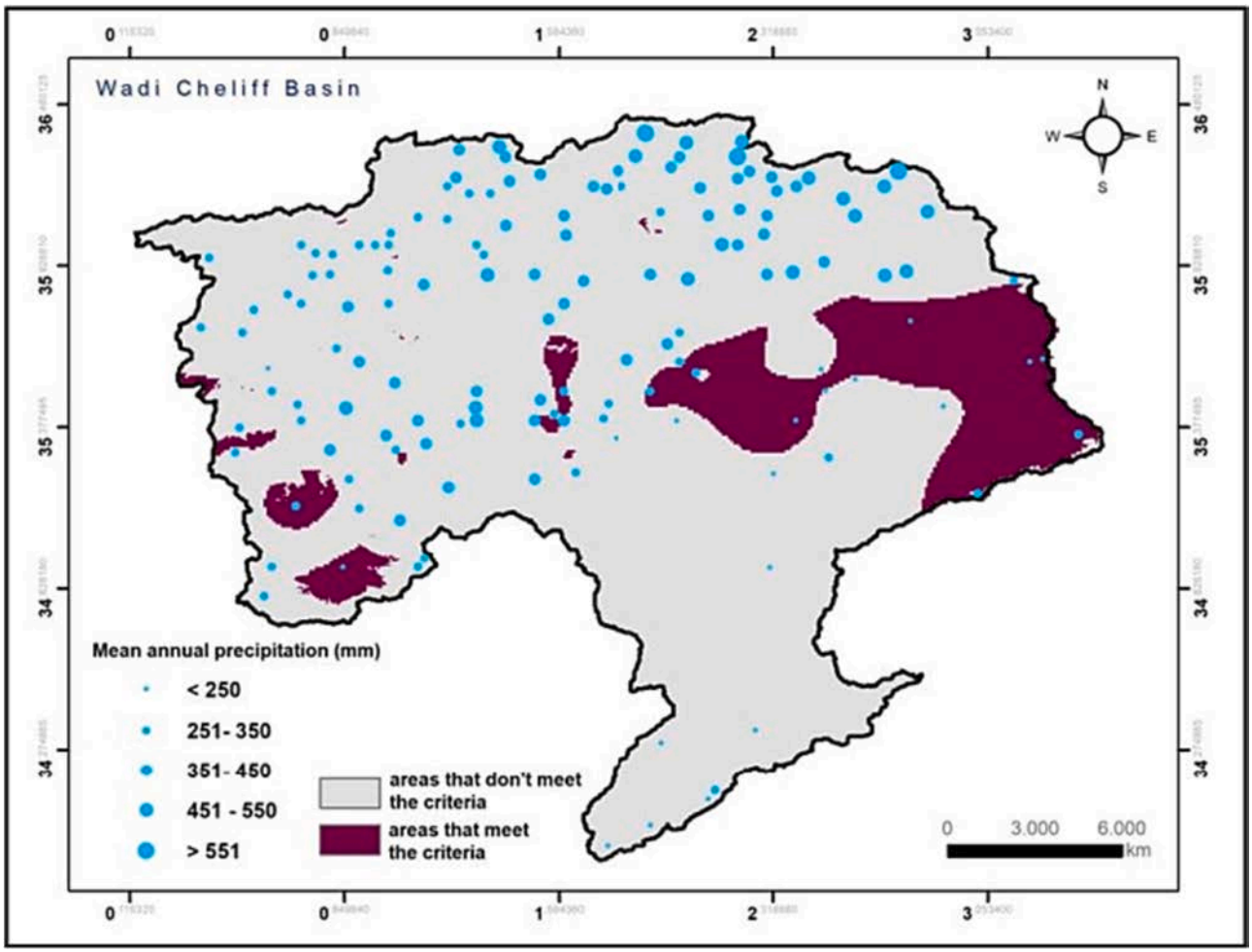

Fig. 10. Spatial distribution of areas with residuals < 15 mm.

group of "weak learners" can formulate a "strong learner" [56]. In more detail, fifty-two (52) and twenty-two (22) stations from the training and test subset, almost half of the entire database, when modeled by the RF model, provided an overestimation compared to the actual precipitation values (Fig. 11). In reference to the elevation zone, the RF model had the best learning r square value in the C elevation zone (600–1,000 m) which reached the value of 0.97. In contrast to the above, the RF model achieved the best predictive r square value (0.96) in areas above 1,000 m (elevation zone D) (Fig. 12). Concerning the cluster analysis, cluster D showed the highest learning performance (r square = 0.94) and also the highest predictive performance (r square = 0.80). Cluster D is characterized by the highest mean values in reference to the parameters "maximum", "minimum", and "SD", the second highest mean precipitation values and the third highest mean CV value. In general, the predictions of the RF model could be characterized as highly local; they appear to be not as susceptible to the influence of outliers as other ML models or statistical methods such as the linear regression [57]. In addition to the favorable features of RF, some limitations should also be mentioned. One of the known disadvantages of RF models is the lack of interpretability. Due to the fact that the final prediction is derived from a forest of trees, it is not feasible to conclude the way predictions are made. Only single trees can be plotted to show the decisions upon which the predictions are based. Furthermore, the process of averaging over all regression trees prevents the ability to make predictions beyond the range of response values in the training data or to predict the entire range of response values [58].

As already mentioned, in terms of learning performance, close to RF model was the MLPNN model. The best learning performance was achieved within the zone elevation zone C (601–1,000 m, R^2 0.94) followed by the elevation zone D (>1,000 m, R^2 0.93) (Fig. 13). Lower performance was observed at the lower elevation zones, whereas the overall R^2 value reached 0.92. Most values are concentrated within 300 and 400 mm. From Fig. 13 there is no evidence of a clear spatial trend concerning overestimation or underestimation. Concerning the cluster analysis the MLPNN model, had the best learning performance in cluster D, as the RF model had.

A different trend was observed when analyzing the test subset, since the predictive performance was much higher in lower altitudes (elevation zone A, less than 200 m) than in higher altitudes (Fig. 14). The overall R^2 falls to 0.71, whereas the same trend has been observed to the other three metrics MAPE, MAE, NSE. Cluster D appears with the best predictive performance (R^2 0.93). The significant difference between the learning and predictive performance of the MLPNN model, could be attributed to the fact that there might be a change of overestimation effect. The tuning process of the structural parameters may require a more sophisticated method to search for the optimal parameters in a wider search area.

Concerning the SVR model, which achieved the second highest predictive performance in terms of the r square value (0.8132) the following could be reported (Fig. 15). The best learning performance was achieved within the elevation zone D (>1,000 m, R^2 0.85) followed by the elevation zone C (601–1,000 m, R^2 0.80). Also, a lower performance was observed at the lower elevation zones, whereas the overall R^2 value reached 0.79. Most values are concentrated within 250 and 460 mm. Within those values an overestimation trend could be identified. Concerning the cluster analysis the SVR model, had the best learning

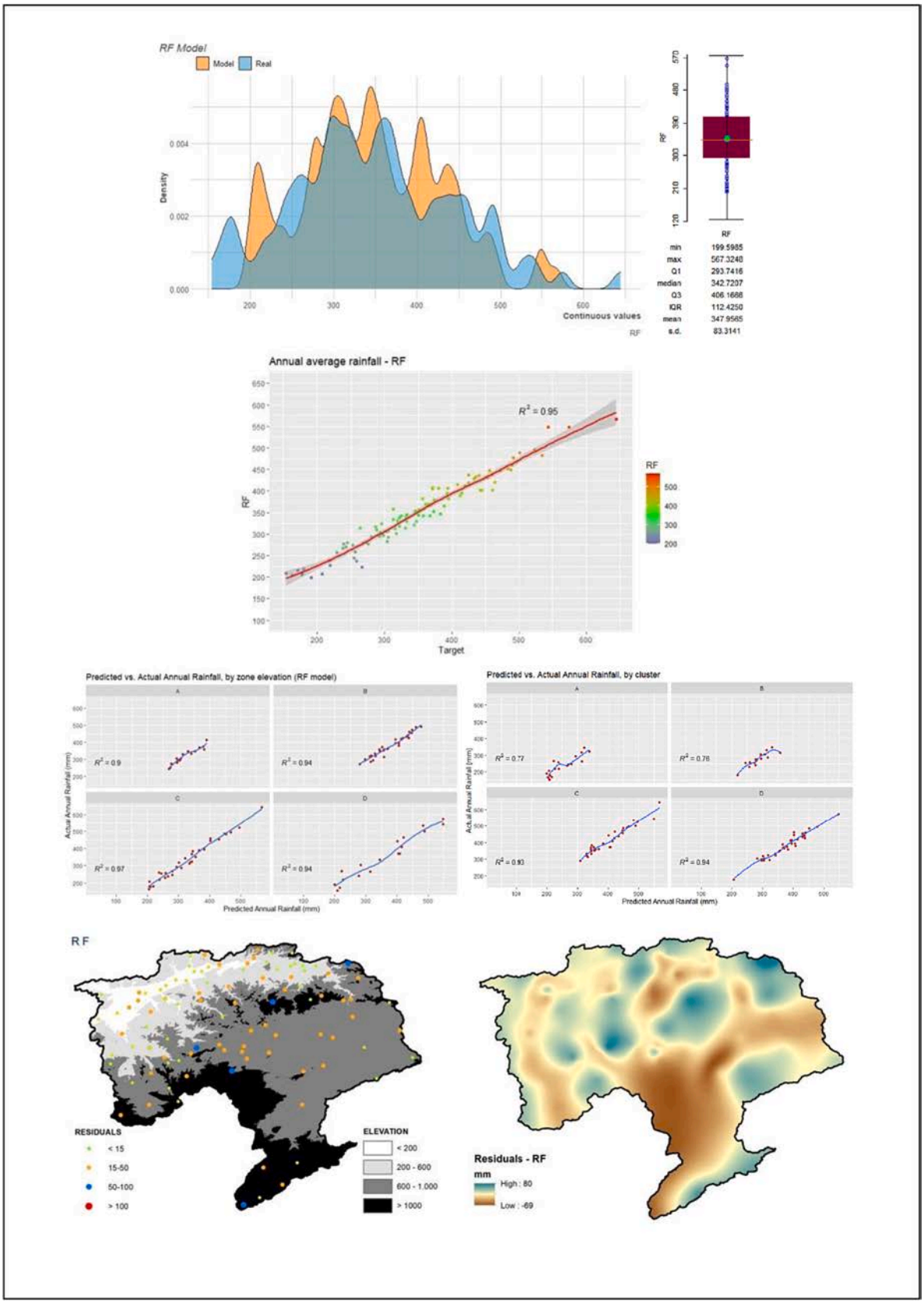

Fig. 11. Information of RF training subset.

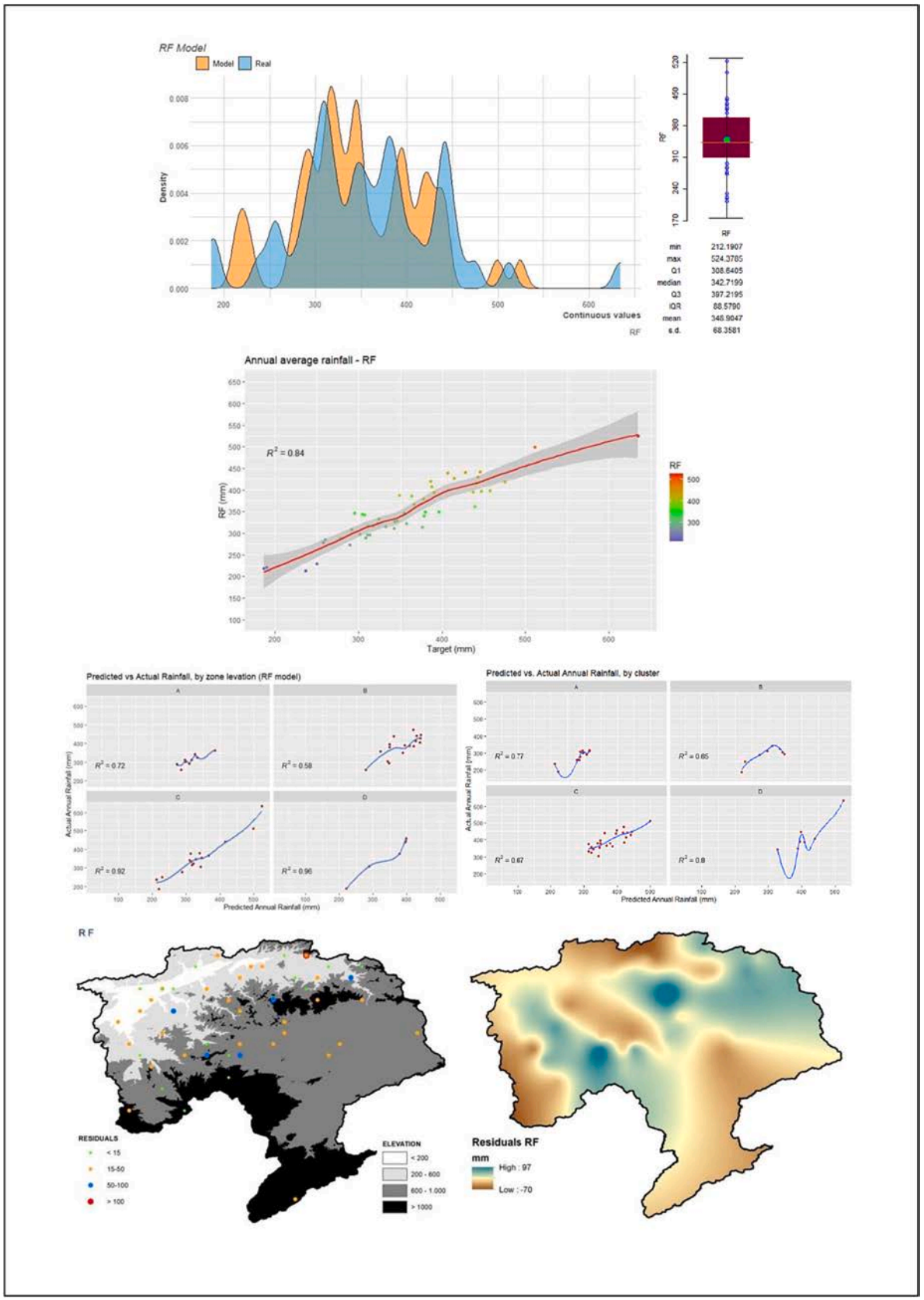

Fig. 12. Information of RF test subset.

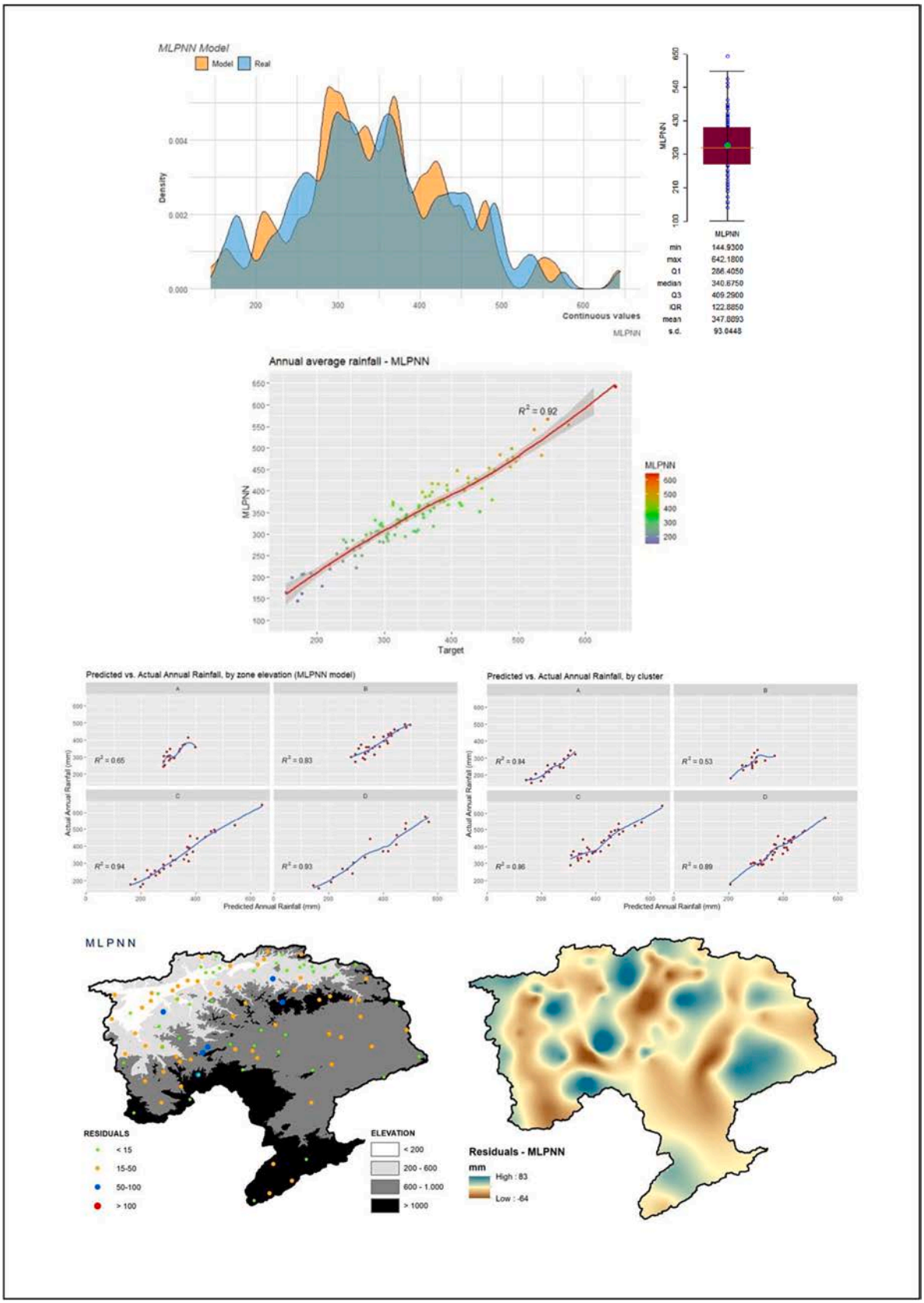

Fig. 13. Information of MLPNN train subset.

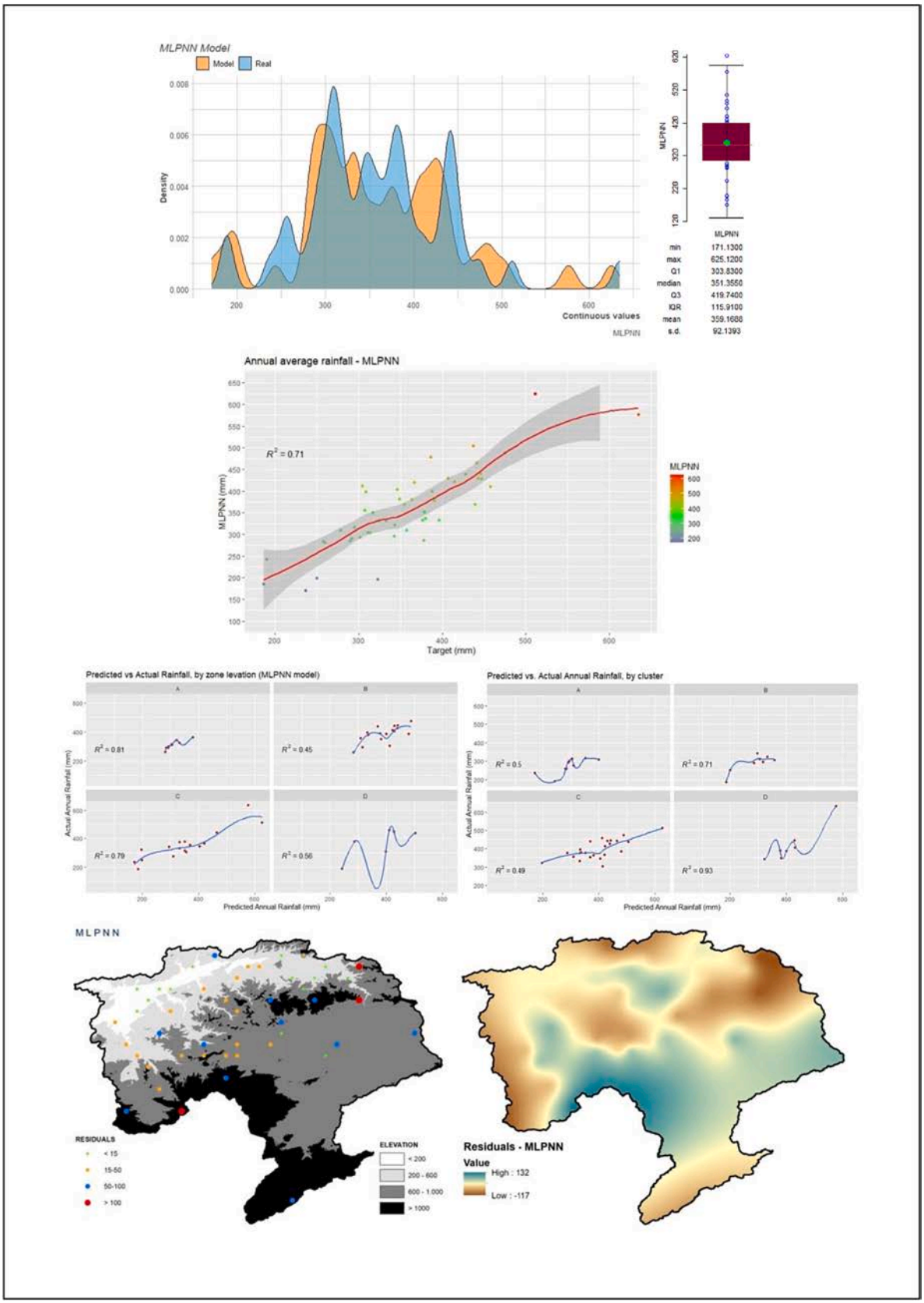

Fig. 14. Information of MLPNN test subset.

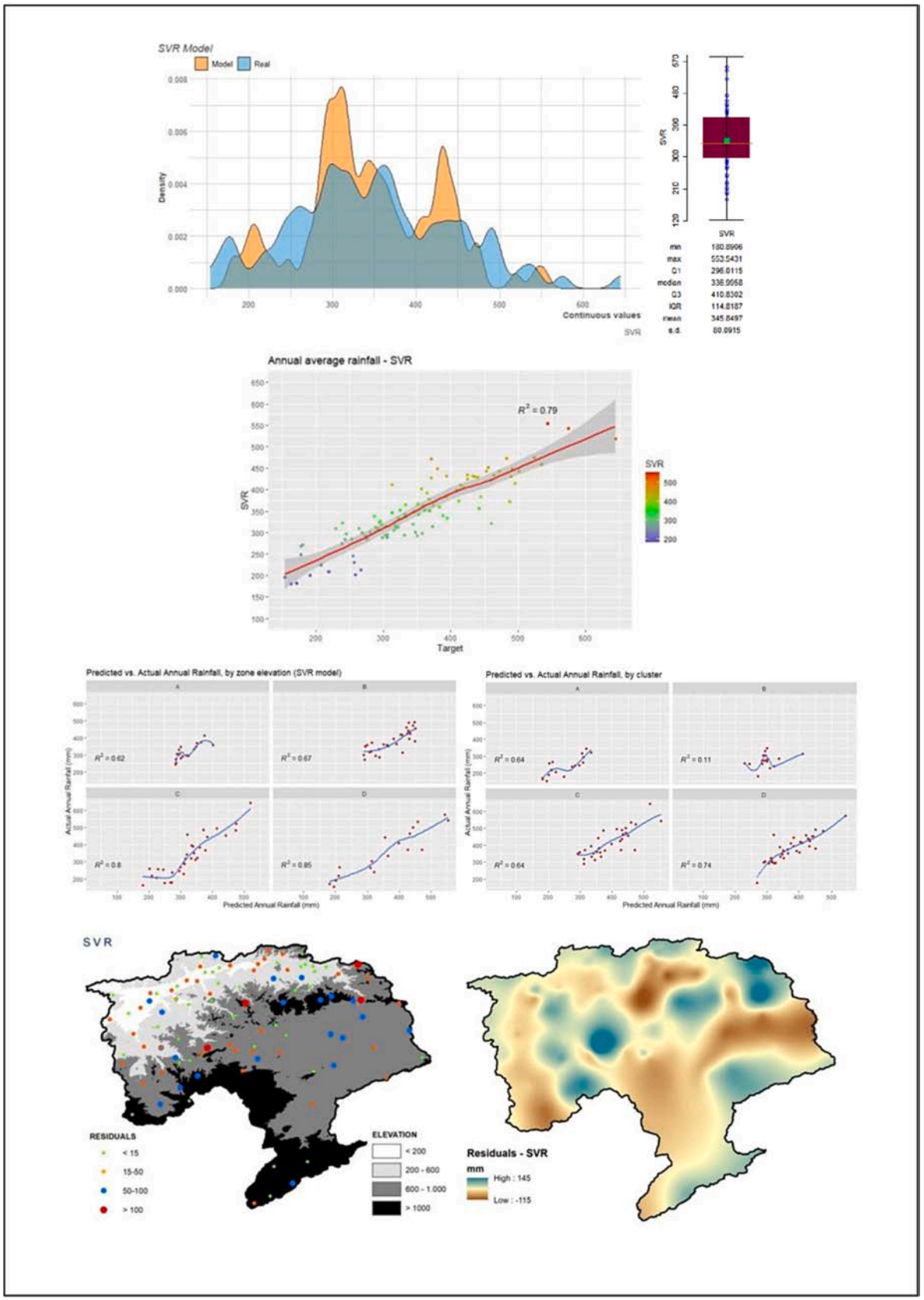

Fig. 15. Information of SVR train subset.

performance in cluster D, as the RF and the MLPNN model had.

A similar trend was observed when analyzing the test subset, since the predictive performance was higher in higher altitudes (Fig. 16). However, in this case the best learning performance was achieved within the elevation zone C (601–1,000 m, R^2 0.90) followed by the elevation zone D (>1,000 m, R^2 0.86), whereas the overall R^2 reaches the value of 0.81. Similar to the training database, cluster D appears with the best predictive performance (R^2 0.88).

Another ML model that has to be analyzed is the M5 model. The M5 model achieved the third highest learning and predictive performance in terms of the r square value (0.7983 and 0.8082) (Fig. 17). The best learning performance was achieved within the zone elevation zone C (601–1000 m, R^2 0.84) followed by the elevation zone D (>1,000 m, R^2 0.80). Also, a lower performance was observed at the lower elevation zones, whereas the overall R^2 value reached 0.80. Most values are concentrated within 250 and 480 mm. There is no evidence of a clear trend concerning an overestimation or underestimation performance. As regards the cluster analysis the M5 model, had the best learning performance in cluster D (0.73).

A quite similar trend was observed when analyzing the test subset (Fig. 18). The best predictive performance was achieved within the elevation zone C (601–1,000 m, R^2 0.89), however slightly lower values have been observed at the elevation zone A (<200 m, R^2 0.87), whereas the overall R^2 reaches the value of 0.81. Similar to the training database, cluster D appears with the best predictive performance (R^2 0.89).

Among the GS models, COK and KED models were the two models that showed the best performance on the training and test subsets and in all the statistical metrics. GS models do not need a training and test phase. The estimation of their performance have been on the two subsets (train and test subset) were made only for comparison with the ML models.

The COK model achieved the highest learning and second predictive performance in terms of the r square value (0.7147 and 0.6255) (Fig. 19). The best learning performance was achieved within the zone elevation zone C (601–1,000 m, R^2 0.75) followed by the elevation zone D (>1,000 m, R^2 0.70). A lower performance was observed at the lower elevation zones, whereas the overall R^2 value reached 0.71. Most values are concentrated within 250 and 450 mm. There is no evidence of a clear trend concerning an overestimation or underestimation performance. Concerning the cluster analysis the COK model, had the best learning performance in cluster D (0.75).

Similarly, with the above, the best learning performance was achieved within the elevation zone C (601–1,000 m, R^2 0.72), whereas the overall R^2 has fallen to the value of 0.63. At the same time, the training database, cluster D appears to have the best predictive performance (R^2 0.70) (Fig. 20).

The KED model achieved the second best learning and highest predictive performance in terms of the r square value (0.6944 and 0.7302) (Fig. 21). The best learning performance was achieved within the zone elevation zone C (601–1000 m, R^2 0.75) followed by the elevation zone D (>1,000 m, R^2 0.67). A lower performance was observed at the lower elevation zones, whereas the overall R^2 value reached 0.69. Most values are concentrated within 250 and 500 mm. There is no evidence of a clear trend concerning an overestimation or underestimation performance. Concerning the cluster analysis the KED model, had by far the best learning performance in cluster D (0.77).

The KED model as already mentioned had the best predictive performance among the GS models, with the overall r square value reaching 0.73 (Fig. 22). The best predictive performance was achieved within the zone elevation zone C (601–1,000 m, R^2 0.83) followed by the elevation zone D (>1,000 m, $R^2$0.78). Concerning the cluster analysis the KED model, had the best learning performance in cluster C (0.61).

Overall, the RF, SVR and M5 models were estimated as the most accurate models in terms of predictive performance, with the RF model achieving in all statistical metrics the best values. Similar results have been obtained from Yaseen et al. (2021), whom studied the capability of different versions of ML models for predicting widely utilized drought indices, specifically the standardized precipitation index, across multiple-month horizons [59]. The authors concluded that the RF model showed the best performance with minimal RMSE value.

However, if one compares the difference between the statistical metrics of the trained and test subsets, the lowest fluctuation was estimated for the M5 model, followed by the SVR model. In other words, when the M5 model uses unknown samples the prediction performance, which is high in both cases, does not differ much from the prediction performance achieved when using known samples. This gives the M5 model an advantage over the RF and SVR models. The M5 model could be more capable of capturing future changes in precipitation trends within the research area, than the rest of the models. In general, in tree-based models, such the RF model, for each data sample, only one local model is used for prediction. However, by this way two adjust inputs could give predicted outputs that may be quite different. M5 models in order to overcome this problem updates the local models through smoothing. According to Pal and Deswal [60], whom calculated evapotranspiration using land surface temperature values extracted from images of MODIS sensors, the main advantage of M5 model tree is that it provides the results in a simple and comprehensible form of regression equations. Concerning the GS models, Spline model appeared to have the lowest fluctuation, in reference to the four statistical metrics. From the above one can assume that the accuracy of Spline model is less impacted by small training subsets than other GS models. As a last point in the present study, one should report the potential similarities among the tested models in reference to Pearson's coefficient as illustrated in Fig. 23. The highest correlation, based on the training subset, was observed among the SVM and QRNN model (0.998), followed by the MLR and QRNN model (0.996) and SVM and MLR model (0.994). Slightly lower coefficient values were observed between the EBK and OK (0.989) and the KED and COK (0.985) models.

A similar trend was observed based on the test subset as illustrated in Fig. 24. The highest correlation was observed again between the SVM and QRNN model (0.997), followed by the MLR and QRNN model (0.995) and SVM and MLR model (0.992). Slightly lower coefficient values were observed between the EBK and OK (0.960) and the EBK and Spline (0.956) models.

The high correlation among the SVM, QRNN and MLR assumes that they perform prediction in a similar linear approach, which however is characterized by low predictive performance. The RF model shows a high correlation with the MLPNN model (0.975), followed by the SVR (0.945) concerning the training subset and a similar high correlation with the SVR (0.945) and RBFNN (0.923) with reference to the test subset.

From the conducted analysis it became apparent that the complexity of the modeling technique plays a pivotal role in performance divergence. ML models, such as RF and Multi-Layer Perceptron Neural Networks (MLPNN), have the capacity to capture intricate nonlinear relationships in the data, which may be advantageous in regions with complex precipitation patterns. Conversely, GS models like Co-Kriging (COK) and Kriging with External Drift (KED) rely on assumptions of spatial autocorrelation and may perform differently in areas where these assumptions are violated. The presence or absence of spatial autocorrelation, which denotes the degree of similarity in precipitation values between neighboring locations, can impact interpolation techniques like Ordinary Kriging (OK) and their performance. Spatial patterns, particularly in regions with complex terrain, can deviate from standard assumptions, affecting the reliability of interpolation methods. Also, the size and distribution of the training dataset are critical factors influencing model performance. In regions with sparse or unevenly distributed gauge stations, models may struggle to capture local variability. Each model could perform in a rather different manner when the size and distribution of the training dataset are changed.

Addressing the critical aspect of uncertainty associated with our models, we adopted a multi-faceted approach to quantify and analyze

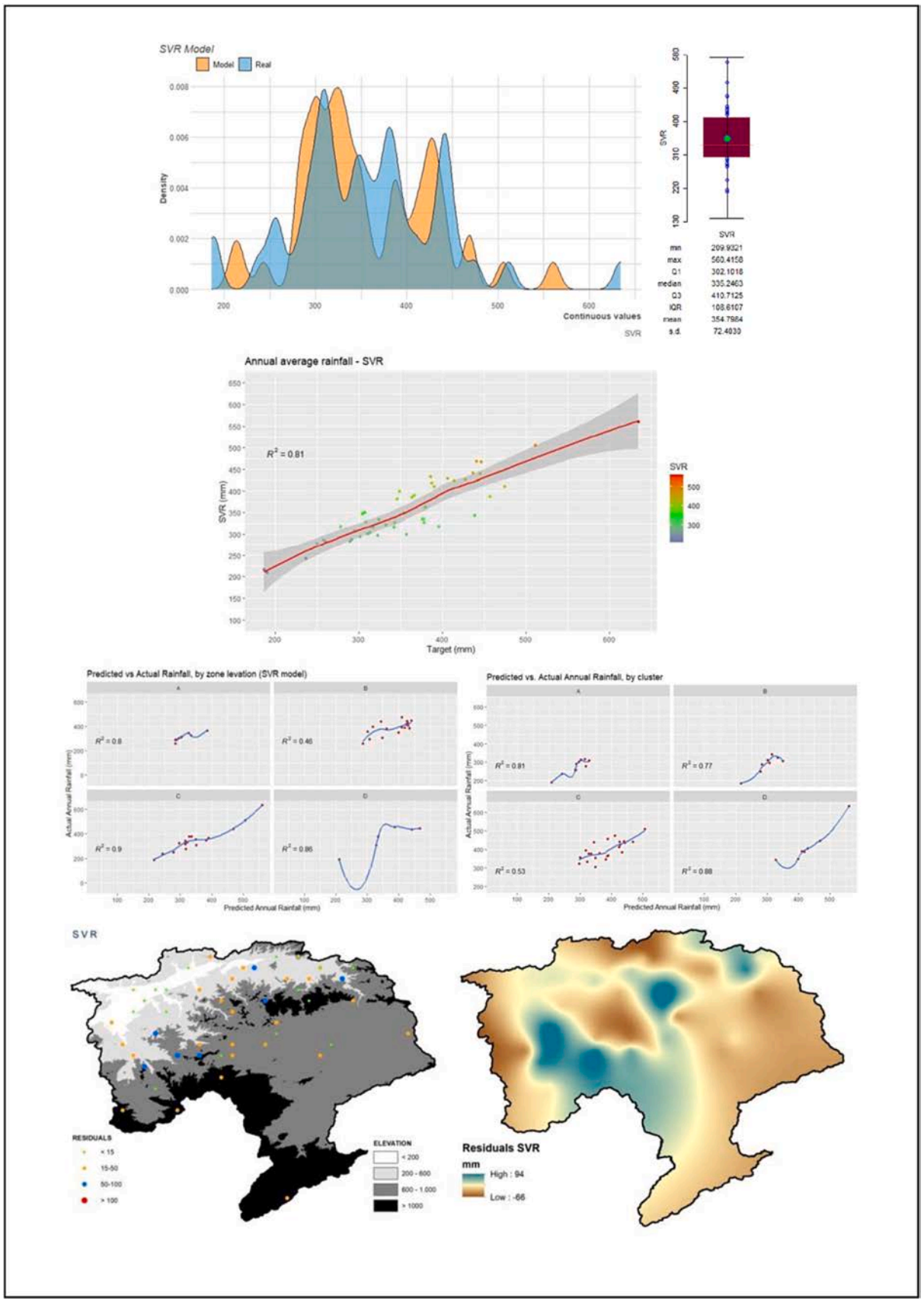

Fig. 16. Information of SVR test subset.

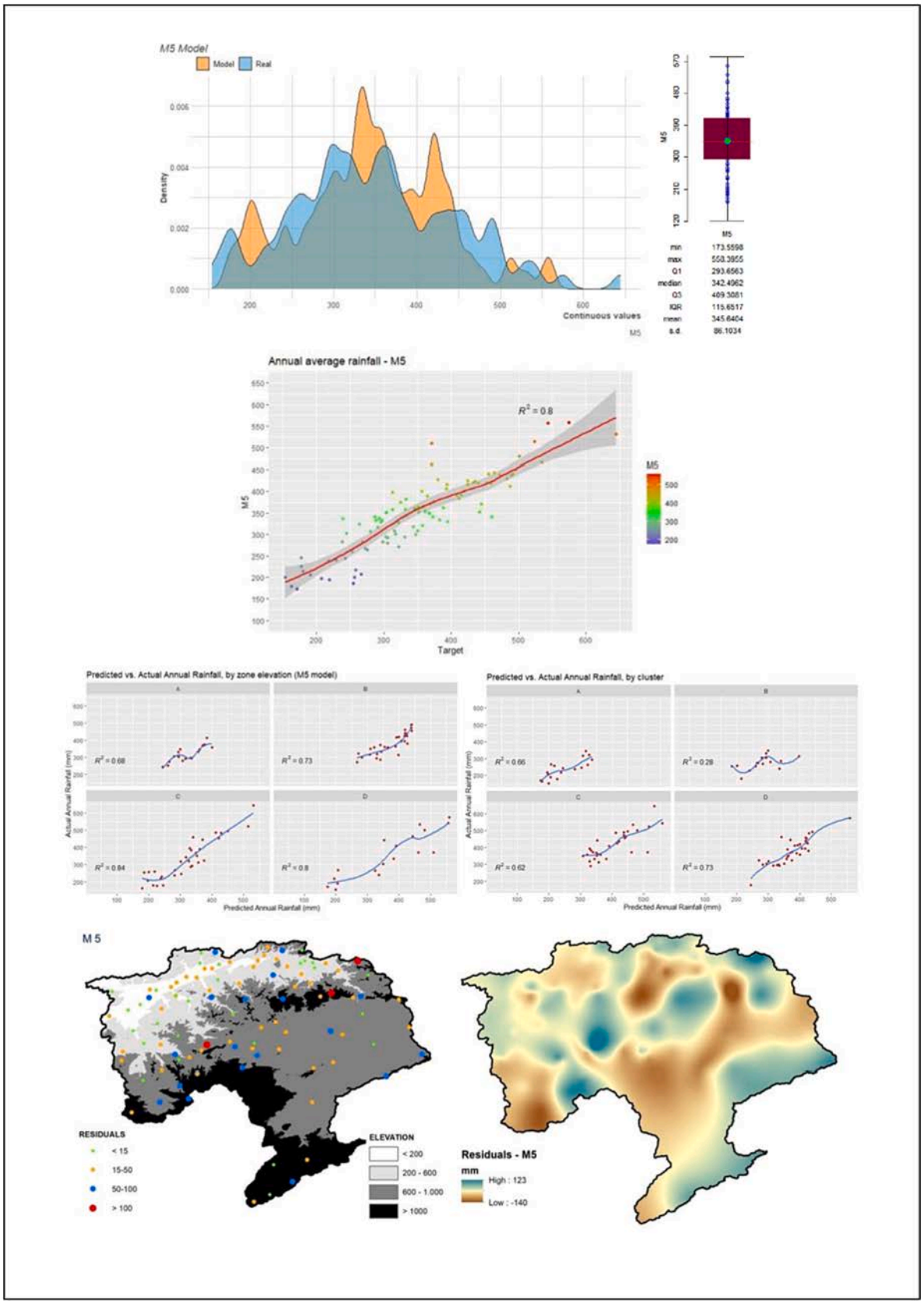

Fig. 17. Information of M5 train subset.

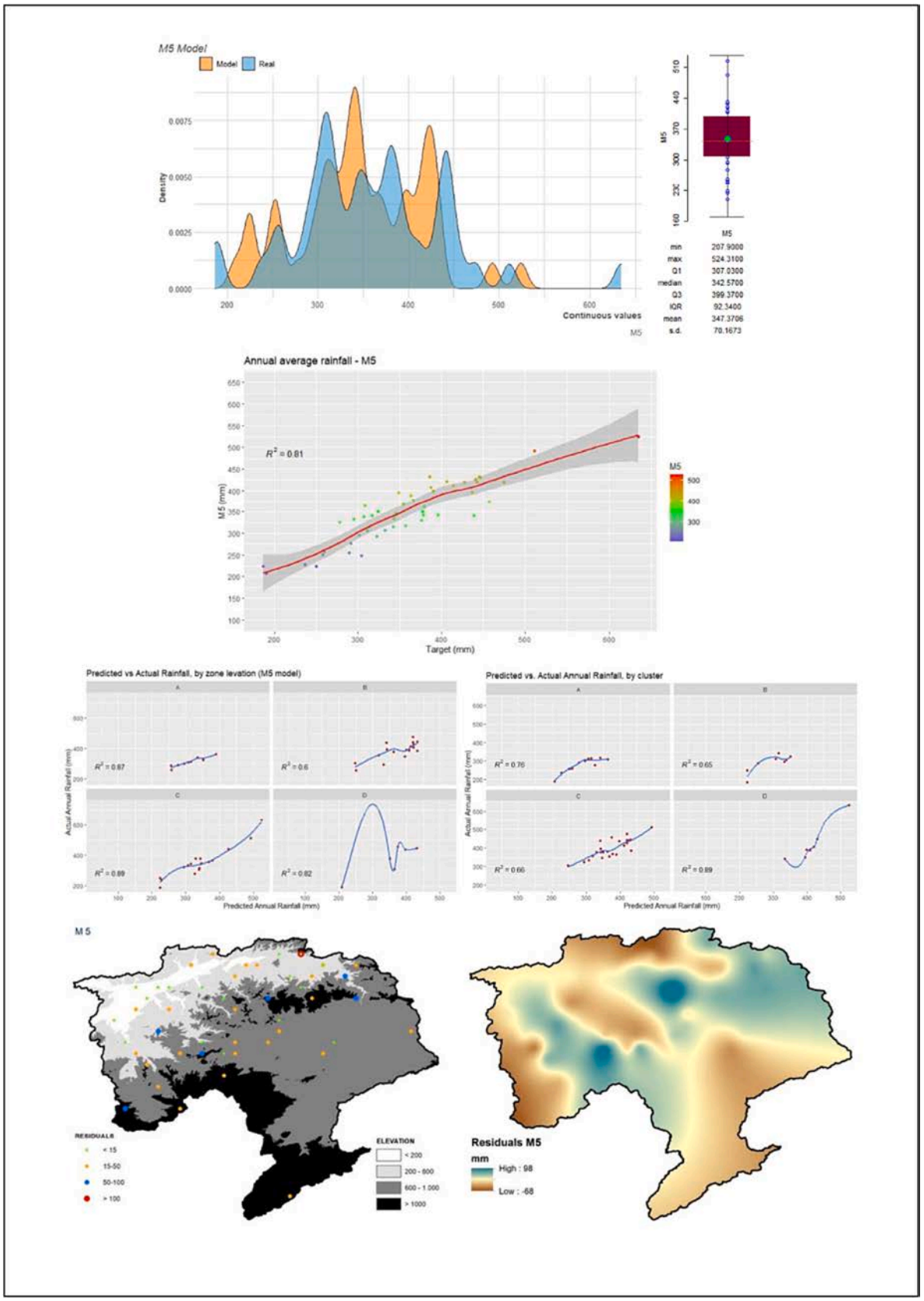

Fig. 18. Information of M5 test subset.

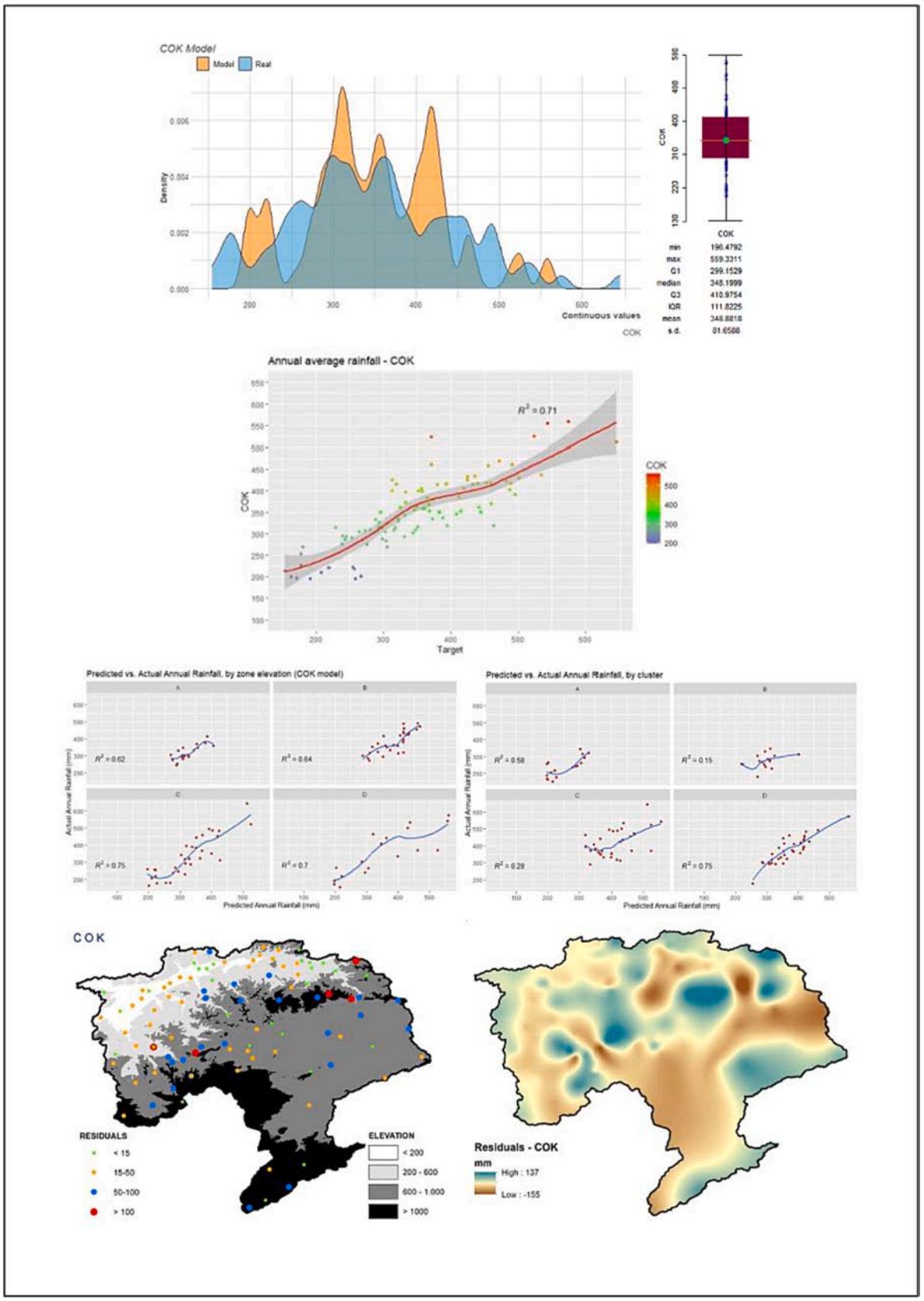

Fig. 19. Information of COK train subset.

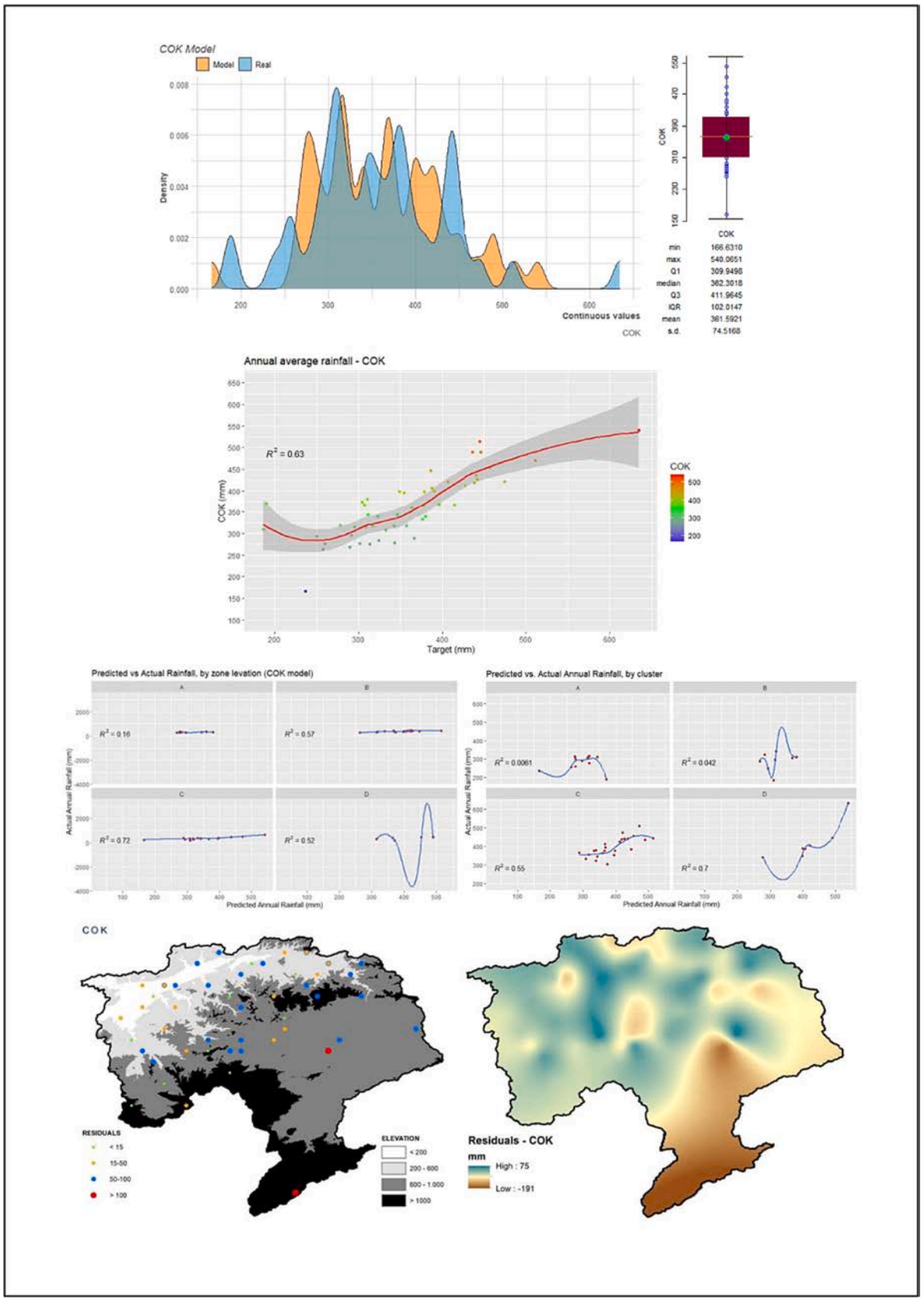

Fig. 20. Information of COK test subset.

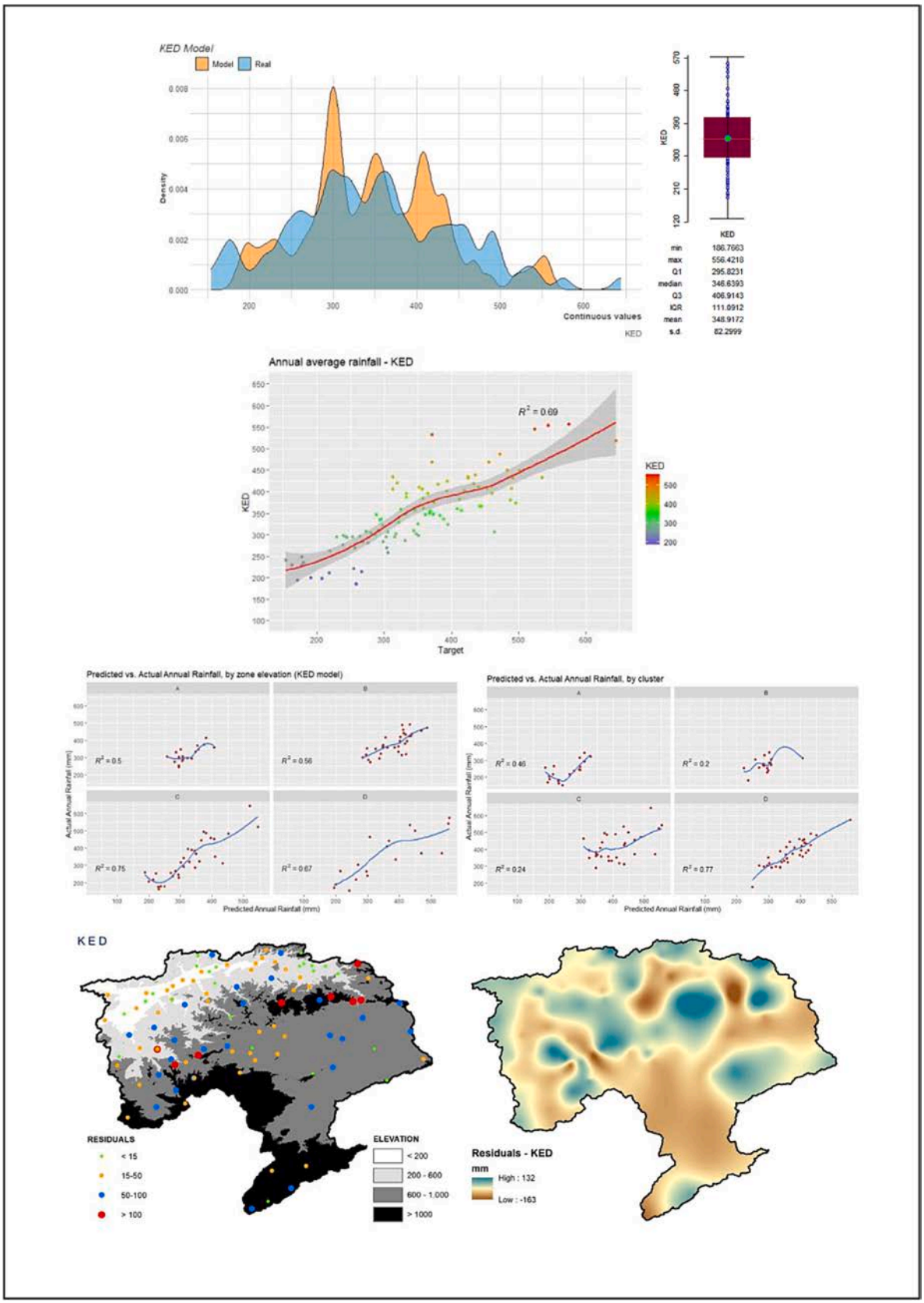

Fig. 21. Information of KED train subset.

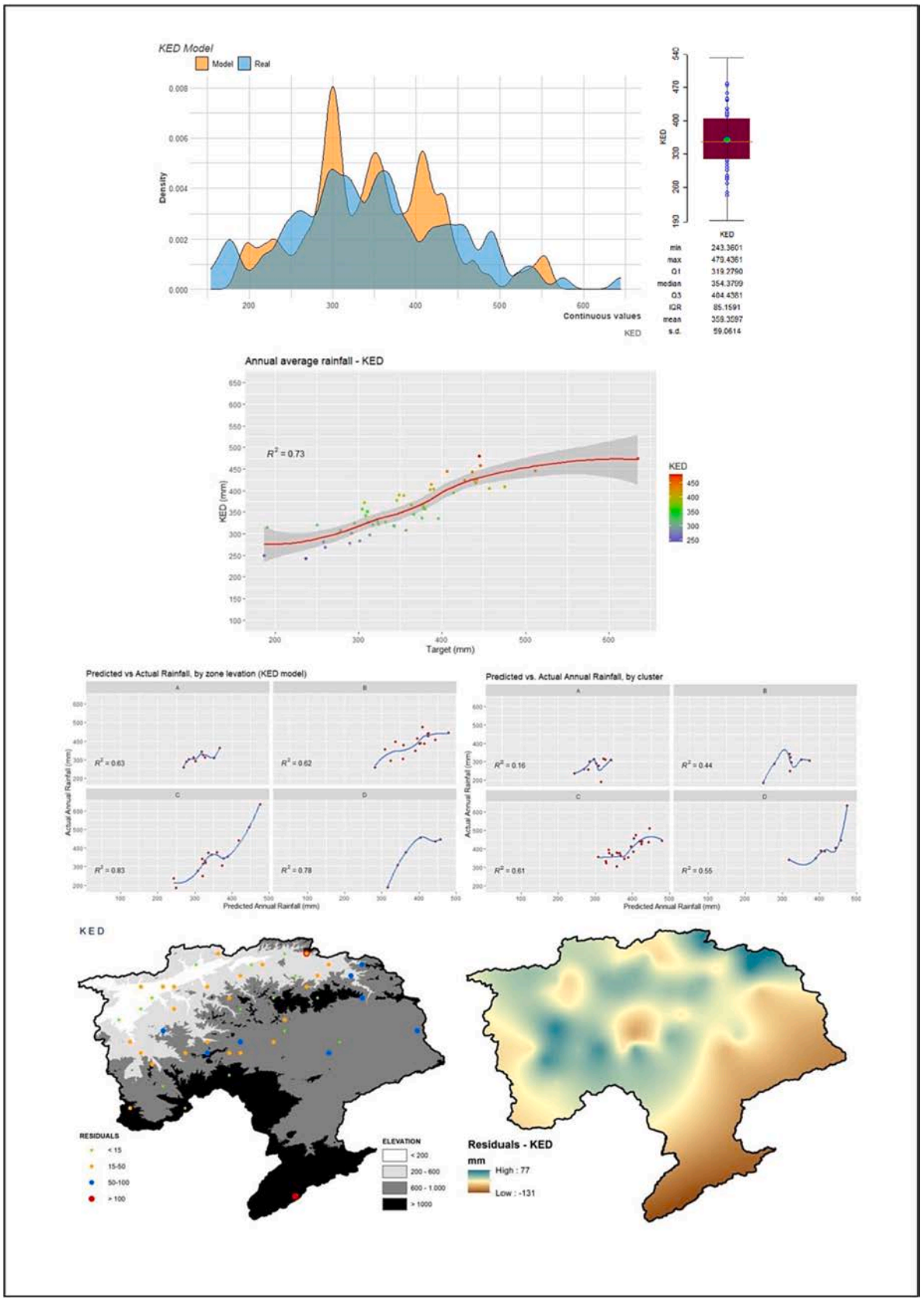

Fig. 22. Information of KED test subset.

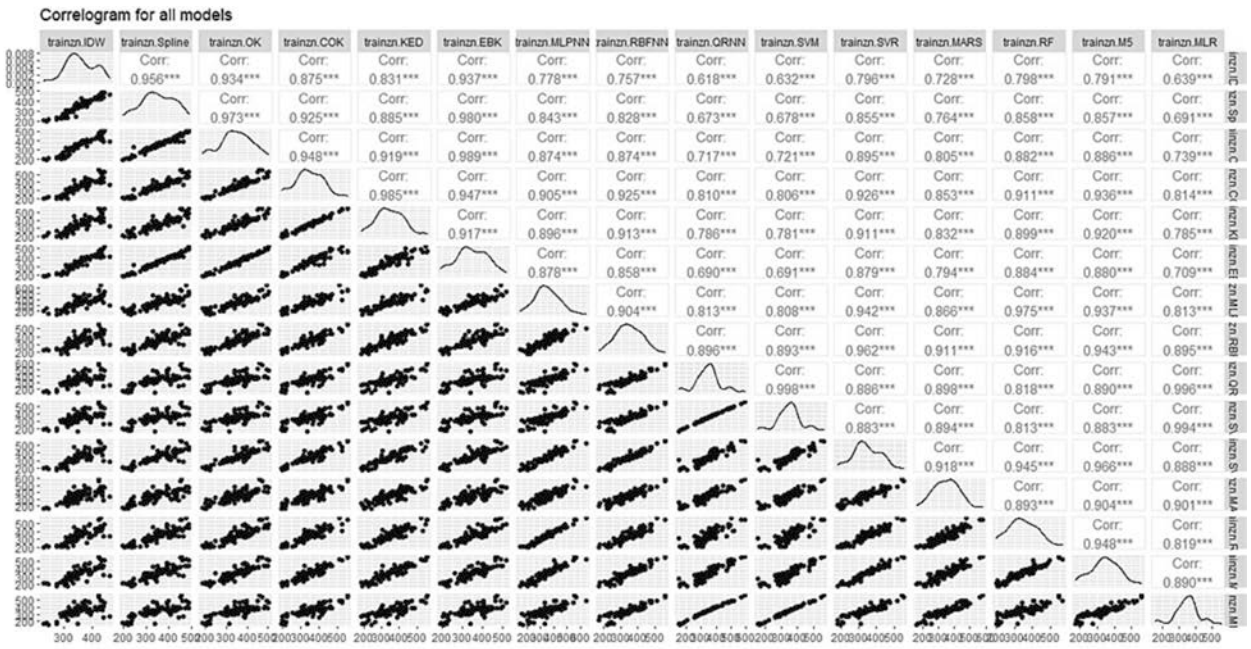

Fig. 23. Correlation matrix of applied models in training subset.

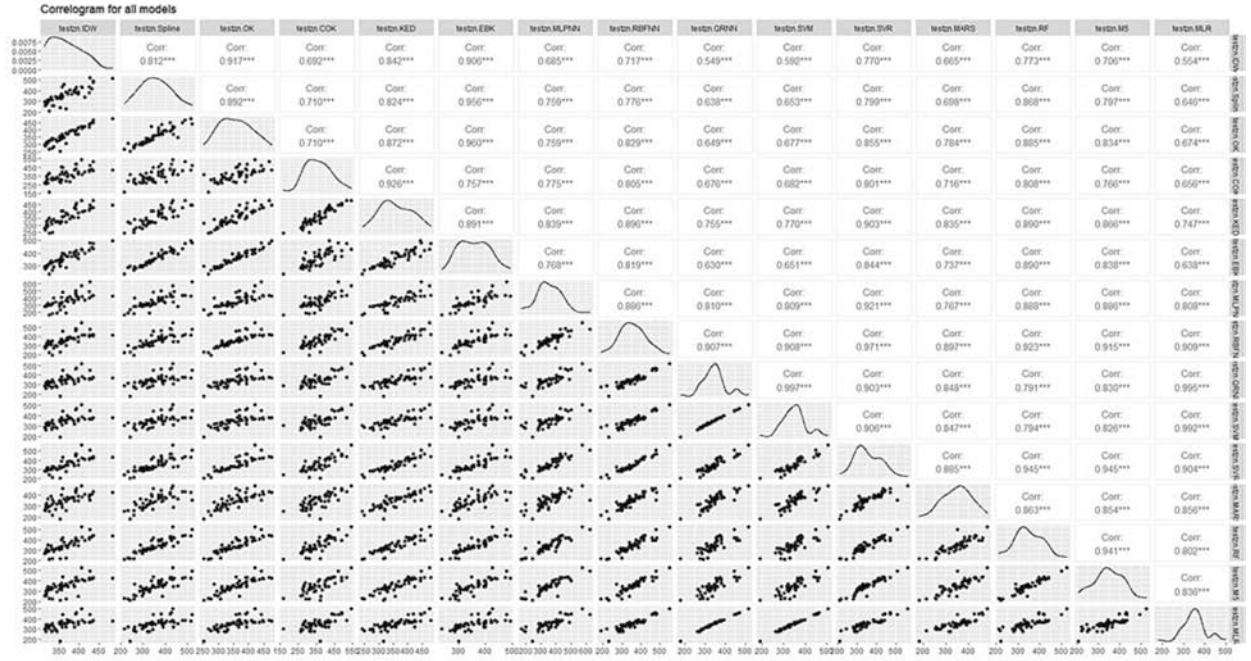

Fig. 24. Correlation matrix of applied models in testing subset.

the sources of variability and error. Residual analysis, a fundamental technique, allowed us to scrutinize the discrepancies between our model's predictions and observed data. By examining the spatial distribution of residuals, we gained valuable insights into areas where our models performed exceptionally well and regions where they exhibited deviations from actual observations. These localized clusters of high residuals hint at potential sources of uncertainty and challenges in capturing complex underlying processes. Additionally, we calculated the Coefficient of Variation (CV) for our model's predictions, providing a measure of relative variability across different locations. Higher CV values signify greater variability and, consequently, potential uncertainty in our predictions. By incorporating these uncertainty measures, we enhance the robustness and reliability of our results, a vital step in supporting informed decision-making in water resource management and climate studies which also could be useful for studies related to the hydrological modeling, environmental modeling, and atmospheric studies [61–63].

Although the analysis conducted has provided some interesting results, one must highlight the limitations. The analysis is characterized of its regional focus and reliance on historical gauge station data, which may exhibit variations in quality and coverage. The selection of modeling approaches, though suitable for the research area and objectives, might not generalize well to different regions with distinct climatic and geographical characteristics. Additionally, the study's models, like

many in the field, operate under specific assumptions about data distributions and relationships, potentially leading to inaccuracies when applied in diverse settings. These limitations highlight the need for cautious interpretation of results and suggest that further research should address these challenges to enhance the models' applicability and reliability in various contexts.

Future research directions may involve expanding the geographic scope by incorporating data from diverse regions with varying climate, topography, and land cover, facilitating a comprehensive assessment of model performance and regional factors influencing predictions. Investigating the transferability of models across different geographic areas will be crucial, as will quantifying and characterizing uncertainties associated with precipitation models, including data quality, model assumptions, and parameterization choices. Additionally, exploring model ensembles, bias correction techniques, and hydrological applications presents promising avenues to advance the field of precipitation modeling, ultimately improving our understanding of hydrology and climate science.

5. Conclusion

Precipitation assessment is a critical aspect of understanding and managing water resources, and it often presents complex challenges due to its spatial and temporal variability. In this study, we conducted a comprehensive analysis of various models, including deterministic, GS, and ML approaches, to predict mean annual precipitation across a diverse geographical area. The spatial interpolation of annual rainfall measured 150 stations in the Wadi Cheliff Basin in the period 1970–2018 was studied using several spatial interpolation techniques: Inverse Distance Weighting, Spline, Ordinary Cokriging, Ordinary Kriging, Empirical Bayesian kriging, Kriging with External Drift, Multilayer Perceptron Neural Networks, Radial Basis Function Neural Network, Quantile Regression Neural Network, Support Vector Machines, Support Vector Regression, Multivariate Adaptive Regression Spline, Random Forest, M5 Model Tree ,and Multiple Linear Regression. The results comparative of the several techniques of spatial interpolation, performed applying the Root Mean Squared Error (RMSE), the Nash-Sutcliffe efficiency (NSE), Mean Absolute Error (MAE), the coefficient of determination (R-squared), and a visual interpretation with the Taylor diagram evidenced acceptable performance of the used models, also as a result of the high-density measure network.

The current study resulted that ML models outperformed GS and deterministic models in predicting mean annual precipitation. Among the ML models, RF demonstrated the highest training and predictive performance, excelling in terms of R-squared values, RMSE, NSE, and MAE. The superiority of the RF model can be attributed to its ensemble-based approach, which harnesses the collective strength of multiple decision trees. In particular, results evidenced that the RF model showed the highest training performance, with a R-square value of 0.9524 and a RMSE value of 24.98. Similarly, the RF model showed the best learning and predictive performance concerning the NSE and the MAE metrics (0.9352 and 19.62 during training, 0.8244, and 26.51 during testing, respectively).

Spatial analysis of the study area uncovered significant variations in mean annual precipitation. Higher precipitation values were concentrated in the northern regions, characterized by lower altitudes, while lower values were observed in the southern and eastern parts. The spatial distribution analysis also highlighted the impact of elevation, with mountainous and complex terrain areas exhibiting enhanced precipitation prediction when elevation was used as an auxiliary variable. Four distinct clusters (A, B, C, and D) were identified, each exhibiting unique precipitation patterns and elevation zones. Notably, cluster C demonstrated the highest mean annual precipitation and cluster D displayed the highest standard deviation, minimum, and maximum values, reflecting the geomorphological diversity of the study area. While the RF model emerged as the top-performing model in both training and test subsets, it is essential to consider the trade-offs between model complexity and interpretability. The M5 model demonstrated remarkable stability between training and test subsets, suggesting its potential for capturing future changes in precipitation trends. The GS models, particularly COK and KED, showcased robust performance and highlighted their utility in areas with limited data availability.

By advancing our understanding of precipitation patterns, this research contributes to more informed environmental management and adaptation strategies in regions characterized by diverse climatic conditions and terrain. To enhance predictions further, researchers could explore the integration of additional variables, such as remote sensing data and climate model outputs. Investigating the impact of climate change on precipitation patterns and extending the analysis to longer temporal scales are promising avenues for future research.

CRediT authorship contribution statement

Mohammed Achite: Conceptualization, Methodology, Software, Validation, Investigation, Writing – original draft, Writing – review & editing, Visualization, Supervision. **Paraskevas Tsangaratos:** Methodology, Software, Validation, Investigation, Writing – original draft, Writing – review & editing, Supervision. **Gaetano Pellicone:** Methodology, Software. **Babak Mohammadi:** Conceptualization, Writing – review & editing, Supervision. **Tommaso Caloiero:** Methodology, Validation, Investigation, Writing – original draft, Writing – review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors appreciate the data collected by the ANRH agency and the Algerian General Directorate of Scientific Research and Technological Development (DGRSDT).

References

[1] Pellicone G, Caloiero T, Guagliardi I. The De Martonne aridity index in Calabria (Southern Italy). J Maps 2019;15. https://doi.org/10.1080/17445647.2019.1673840.

[2] Bui DT, Pradhan B, Revhaug I, Nguyen DB, Pham HV, Bui QN. A novel hybrid evidential belief function-based fuzzy logic model in spatial prediction of rainfall-induced shallow landslides in the Lang Son city area (Vietnam). http://wwwTandfonlineCom/Action/JournalInformation?Show=aimsScope&journalCode=tgnh20#VsXodSCLRhE 2013;6:243–71. https://doi.org/10.1080/19475705.2013.843206.

[3] Mohammadi B, Vaheddoost B, Danandeh MA. A spatiotemporal teleconnection study between Peruvian precipitation and oceanic oscillations. Quat Int 2020. https://doi.org/10.1016/j.quaint.2020.09.042.

[4] Goovaerts P. Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. J Hydrol (amst) 2000;228. https://doi.org/10.1016/S0022-1694(00)00144-X.

[5] Chu KS, Oh CH, Choi JR, Kim BS. Estimation of threshold rainfall in ungauged areas using machine learning. Water. 2022;Vol 14:859. https://doi.org/10.3390/W14060859.

[6] Guo B, Luo W, Han B, Liu Y, Yang F, Fan Y, et al. Multiple spatial–temporal scale change patterns of rainfall erosivity in China over past 58 years based on gravity centre model. Http://WwwTandfonlineCom/Action/JournalInformation?Show=aimsScope&journalCode=tgnh20#VsXodSCLRhE 2019;10:2200–19. https://doi.org/10.1080/19475705.2019.1685012.

[7] Buttafuoco G, Caloiero T, Ricca N, Guagliardi I. Assessment of drought and its uncertainty in a southern Italy area (Calabria region). Measurement (lond) 2018;113. https://doi.org/10.1016/j.measurement.2017.08.007.

[8] Delbari M, Afrasiab P, Jahani S. Spatial interpolation of monthly and annual rainfall in northeast of Iran. Meteorol Atmos Phys 2013;122. https://doi.org/10.1007/s00703-013-0273-5.

[9] Li J, Heap AD. A review of comparative studies of spatial interpolation methods in environmental sciences: Performance and impact factors. Ecol Inform 2011;6. https://doi.org/10.1016/j.ecoinf.2010.12.003.

[10] Caloiero T, Pellicone G, Modica G, Guagliardi I. Comparative analysis of different spatial interpolation methods applied to monthly rainfall as support for landscape management. Applied Sciences (Switzerland). 2021;11. https://doi.org/10.3390/app11209566.

[11] di Piazza A, Conti F lo, Noto LV, Viola F, la Loggia G. Comparative analysis of different techniques for spatial interpolation of rainfall data to create a serially complete monthly time series of precipitation for Sicily, Italy. Int J Appl Earth Obs Geoinf 2011;13. https://doi.org/10.1016/j.jag.2011.01.005.

[12] Pellicone G, Caloiero T, Modica G, Guagliardi I. Application of several spatial interpolation techniques to monthly rainfall data in the Calabria region (southern Italy). Int J Climatol 2018;38. https://doi.org/10.1002/joc.5525.

[13] Keblouti M, Ouerdachi L, Boutaghane H. Spatial interpolation of annual precipitation in Annaba- Algeria - Comparison and evaluation of methods. Energy Procedia 2012;18. https://doi.org/10.1016/j.egypro.2012.05.058.

[14] Bachir H, Semar A, Mazari A. Statistical and geostatistical analysis related to geographical parameters for spatial and temporal representation of rainfall in semi-arid environments: the case of Algeria. Arab J Geosci 2016;9. https://doi.org/10.1007/s12517-016-2505-8.

[15] Rata M, Douaoui A, Larid M, Douaik A. Comparison of geostatistical interpolation methods to map annual rainfall in the Chéliff watershed, Algeria. Theor Appl Climatol. 2020:141. https://doi.org/10.1007/s00704-020-03218-z.

[16] Achite M, Caloiero T, Wałęga A, Krakauer N, Hartani T. Analysis of the spatiotemporal annual rainfall variability in the wadi cheliff basin (Algeria) over the period to 2018. Water (Switzerland). 1970;2021:13. https://doi.org/10.3390/w13111477.

[17] Elmeddahi Y, Mahmoudi H, Issaadi A, Goosen MFA, Ragab R. Evaluating the effects of climate change and variability on water resources: A case study of the Cheliff basin in Algeria. American Journal of Engineering and Applied Sciences 2016;9. https://doi.org/10.3844/ajeassp.2016.835.845.

[18] Kuhn M. Building predictive models in R using the caret package. J Stat Softw 2008;28. https://doi.org/10.18637/jss.v028.i05.

[19] Skjelbred HI, Kong J. A comparison of linear interpolation and spline interpolation for turbine efficiency curves in short-term hydropower scheduling problems. IOP Conf Ser Earth Environ Sci, vol. 240, 2019. https://doi.org/10.1088/1755-1315/240/4/042011.

[20] Harder RL, Desmarais RN. Interpolation using surface splines. Https://DoiOrg/102514/344330 2012;9:189–91. https://doi.org/10.2514/3.44330.

[21] Pilz J, Spöck G. Why do we need and how should we implement Bayesian kriging methods. Stoch Env Res Risk A 2008;22. https://doi.org/10.1007/s00477-007-0165-7.

[22] Valentine DL, Fisher GB, Bagby SC, Nelson RK, Reddy CM, Sylva SP, et al. Fallout plume of submerged oil from Deepwater Horizon. PNAS 2014;111. https://doi.org/10.1073/pnas.1414873111.

[23] Krivoruchko K. Empirical bayesian kriging. ESRI Press 2012 Fall 2012.

[24] Krivoruchko K. Spatial statistical data analysis for GIS users spatial statistical data analysis for GIS users. Analysis 2011.

[25] Nanda SK, Tripathy DP, Nayak SK, Mohapatra S. Prediction of Rainfall in india using artificial neural network (ANN) models. International Journal of Intelligent Systems and Applications 2013;5. https://doi.org/10.5815/ijisa.2013.12.01.

[26] Imseng D, Motlicek P, Garner PN, Bourlard H. Impact of deep MLP architecture on different acoustic modeling techniques for under-resourced speech recognition. 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2013 - Proceedings, 2013. https://doi.org/10.1109/ASRU.2013.6707752.

[27] Kwok TY, Yeung DY. Constructive algorithms for structure learning in feedforward neural networks for regression problems. IEEE Trans Neural Netw 1997;8. https://doi.org/10.1109/72.572102.

[28] Mohammadi B, Mehdizadeh S, Ahmadi F, Lien NTT, Linh NTT, Pham QB. Developing hybrid time series and artificial intelligence models for estimating air temperatures. Stoch Env Res Risk A 2021;35. https://doi.org/10.1007/s00477-020-01898-7.

[29] Hong H, Tsangaratos P, Ilia I, Loupasakis C, Wang Y. Introducing a novel multi-layer perceptron network based on stochastic gradient descent optimized by a meta-heuristic algorithm for landslide susceptibility mapping. Sci Total Environ 2020:742. https://doi.org/10.1016/j.scitotenv.2020.140549.

[30] Luo X, Xu Y, Xu J. Application of radial basis function network for spatial precipitation interpolation. 2010 18th International Conference on Geoinformatics, Geoinformatics. 2010, 2010. Doi: 10.1109/GEOINFORMATICS.2010.5567841.

[31] Taylor JW. A quantile regression neural network approach to estimating the conditional density of multiperiod returns. J Forecast. 2000;19. https://doi.org/10.1002/1099-131x(200007)19:4<299::aid-for775>3.3.co;2-m.

[32] Cannon AJ. Non-crossing nonlinear regression quantiles by monotone composite quantile regression neural network, with application to rainfall extremes. Stoch Env Res Risk A 2018:32. https://doi.org/10.1007/s00477-018-1573-6.

[33] Cannon AJ. Quantile regression neural networks: Implementation in R and application to precipitation downscaling. Comput Geosci 2011:37. https://doi.org/10.1016/j.cageo.2010.07.005.

[34] Vapnik VN. An overview of statistical learning theory. IEEE Trans Neural Netw 1999:10. https://doi.org/10.1109/72.788640.

[35] Yao X, Tham LG, Dai FC. Landslide susceptibility mapping based on Support Vector Machine: A case study on natural slopes of Hong Kong, China. Geomorphology 2008:101. https://doi.org/10.1016/j.geomorph.2008.02.011.

[36] Vapnik V, Golowich SE, Smola A. Support vector method for function approximation, regression estimation, and signal processing. Adv Neural Inf Process Syst 1997.

[37] Williams CKI. Learning with kernels: Support vector machines, regularization, optimization, and beyond. J Am Stat Assoc 2003:98. https://doi.org/10.1198/jasa.2003.s269.

[38] Mohammadi B, Mehdizadeh S. Modeling daily reference evapotranspiration via a novel approach based on support vector regression coupled with whale optimization algorithm. Agric Water Manag 2020:237. https://doi.org/10.1016/j.agwat.2020.106145.

[39] Friedman JH. Multivariate adaptive regression splines. Https://DoiOrg/101214/Aos/1176347963 1991;19:1–67. https://doi.org/10.1214/AOS/1176347963.

[40] Sharda VN, Patel RM, Prasher SO, Ojasvi PR, Prakash C. Modeling runoff from middle Himalayan watersheds employing artificial intelligence techniques. Agric Water Manag 2006:83. https://doi.org/10.1016/j.agwat.2006.01.003.

[41] Wolfensberger D, Gabella M, Boscacci M, Germann U, Berne A. RainForest: a random forest algorithm for quantitative precipitation estimation over Switzerland. Atmos Meas Tech 2021:14. https://doi.org/10.5194/amt-14-3169-2021.

[42] Mohammadi B, Guan Y, Moazenzadeh R, Safari MJS. Implementation of hybrid particle swarm optimization-differential evolution algorithms coupled with multi-layer perceptron for suspended sediment load estimation. Catena (amst) 2021;198:105024. https://doi.org/10.1016/j.catena.2020.105024.

[43] Breiman L. Bagging predictors. Mach Learn 1996:24. https://doi.org/10.1007/bf00058655.

[44] Abdallah M, Mohammadi BH, Zaroug MA, Omer A, Cheraghalizadeh M, Eldow MEE, et al. Reference evapotranspiration estimation in hyper-arid regions via D-vine copula based-quantile regression and comparison with empirical approaches and machine learning models. J Hydrol: Reg Stud 2022;44:101259. https://doi.org/10.1016/J.EJRH.2022.101259.

[45] Tien Bui D, Pradhan B, Lofman O, Revhaug I. Landslide susceptibility assessment in vietnam using support vector machines, decision tree, and nave bayes models. Math Probl Eng 2012. https://doi.org/10.1155/2012/974638.

[46] Quinlan JR. Learning with continuous classes. Australian Joint Conference on Artificial Intelligence, 1992.

[47] Rahimikhoob A, Asadi M, Mashal M. A comparison between conventional and M5 model tree methods for converting pan evaporation to reference evapotranspiration for semi-arid region. Water Resour Manag 2013:27. https://doi.org/10.1007/s11269-013-0440-y.

[48] Solomatine DP, Xue Y. M5 Model trees and neural networks: Application to flood forecasting in the upper reach of the huai river in china. J Hydrol Eng. 2004;9. https://doi.org/10.1061/(asce)1084–0699(2004)9: 6 (491).

[49] Legates DR, McCabe GJ. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resour Res 1999:35.

[50] Bostan PA, Heuvelink GBM, Akyurek SZ. Comparison of regression and kriging techniques for mapping the average annual precipitation of Turkey. Int J Appl Earth Obs Geoinf 2012:19. https://doi.org/10.1016/j.jag.2012.04.010.

[51] Meng Q, Liu Z, Borders BE. Assessment of regression kriging for spatial interpolation - Comparisons of seven GIS interpolation methods. Cartogr Geogr Information Scientist 2013:40. https://doi.org/10.1080/15230406.2013.762138.

[52] de Borges PdeA, Franke J, da Anunciação YMT, Weiss H, Bernhofer C. Comparison of spatial interpolation methods for the estimation of precipitation distribution in Distrito Federal, Brazil. Theor Appl Climatol 2016:123. https://doi.org/10.1007/s00704-014-1359-9.

[53] Taylor KE. Summarizing multiple in a single diagram. J Geophys Res 2001;106.

[54] Syakur MA, Khotimah BK, Rochman EMS, Satoto BD. Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster. IOP Conf Ser Mater Sci Eng, vol. 336, 2018. https://doi.org/10.1088/1757-899X/336/1/012017.

[55] Adhikary SK, Muttil N, Yilmaz AG. Cokriging for enhanced spatial interpolation of rainfall in two Australian catchments. Hydrol Process 2017:31. https://doi.org/10.1002/hyp.11163.

[56] Dietterichl TG. The handbook of brain theory and neural networks - ensemble learning. MIT Press; 2002. p. 40.

[57] Sage AJ. Random Forest Robustness, Variable Importance, and Tree Aggregation. 2018.

[58] Kühnlein M, Appelhans T, Thies B, Nauß T. Precipitation estimates from MSG SEVIRI daytime, nighttime, and twilight data with random forests. J Appl Meteorol Climatol 2014:53. https://doi.org/10.1175/JAMC-D-14-0082.1.

[59] Yaseen ZM, Ali M, Sharafati A, Al-Ansari N, Shahid S. Forecasting standardized precipitation index using data intelligence models: regional investigation of Bangladesh. Sci Rep 2021:11. https://doi.org/10.1038/s41598-021-82977-9.

[60] Pal M, Deswal S. M5 model tree based modelling of reference evapotranspiration. Hydrol Process 2009:23. https://doi.org/10.1002/hyp.7266.

[61] Talebmorad H, Ostad-Ali-Askari K. Hydro geo-sphere integrated hydrologic model in modeling of wide basins. Sustain Water Resour Manag 2022;8:1–17. https://doi.org/10.1007/S40899-022-00689-Y/TABLES/9.

[62] Javadinejad S, Eslamian S, Ostad-Ali-Askari K. Investigation of monthly and seasonal changes of methane gas with respect to climate change using satellite data. Appl Water Sci 2019;9:1–8. https://doi.org/10.1007/S13201-019-1067-9/FIGURES/13.

[63] Ostad-Ali-Askar K, Su R, Liu L. Journal of Water and Climate Change | 09 2018;2. https://doi.org/10.2166/wcc.2018.999.