Journal of Hydrology 586 (2020) 124789

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Long-term rainfall prediction using atmospheric synoptic patterns in semi-arid climates with statistical and machine learning methods

Javier Diez-Sierra, Manuel del Jesus*

IHCantabria - Instituto de Hidráulica Ambiental de la Universidad de Cantabria, Avda. Isabel Torres, 15, Parque Científico y Tecnológico de Cantabria, 39011 Santander, Spain

ARTICLE INFO

This manuscript was handled by A. Bardossy, Editor-in-Chief, with the assistance of Uwe Haberlandt, Associate Editor

Keywords:
Rainfall prediction
Regression
Machine learning
Extremes
Statistical downscaling

ABSTRACT

In this paper, we evaluate the performance of 8 statistical and machine learning methods, driven by atmospheric synoptic patterns, for long-term daily rainfall prediction in a semi-arid climate (Tenerife, Spain). Cross-validation is used to reconstruct 36 years of daily rainfall data at 17 gauges. Prediction is independent for each gauge. The reconstructed series are compared with the observed records in order to select the optimal hyperparameters within each family of models. The predictive performance of the models is evaluated using several metrics and statistics related with rainfall intensity and occurrence at daily, monthly and annual aggregation scales. Multivariate and univariate analysis of variance are used to evaluate the differences among models.

The results of our work demonstrate that the performance of most machine learning models is very sensitive to the selected hyperparameters. Neural networks are found to perform best to predict rainfall occurrence and intensity. All methods underestimate the variance of the observed series at daily time scales. Generalized linear models using gamma-distributed errors perform best for predicting rainfall extremes, however, their performance limits its practical applications. Results improve significatively at larger temporal aggregations (monthly or annual) making statistical and machine learning methods more valuable for water resources studies.

1. Introduction

Today, the most reliable measure of the intensity of precipitation at a specific point continues to be that obtained from meteorological stations located on the ground -i.e. instrumental data- (Adeyewa and Nakamura, 2003; Hasan et al., 2016; Wang et al., 2019). This kind of information with a sufficient temporal coverage -usually of several decades- is not commonly available in most areas of the world (Buytaert et al., 2012). However, on-site rainfall information with such a temporal coverage is essential in many fields, such as water resources management, flood risk assessment, and climatological analysis – for instance to detect possible climatic trends in rainfall series-, among others (Nkiaka et al., 2017; Sun et al., 2018; Altunkaynak and Nigussie, 2015; Pumo et al., 2017; Aryal, 2018; Park et al., 2019).

Historical rainfall data, to complement instrumental records, can also be obtain from other sources such as satellites (Huffman et al., 2010), radars (Austin and Seed, 2005) and numerical models (Pfeifroth et al., 2013). Satellites and radars provide short precipitation time series that start, in the best case scenario, in the 1990s (Pfeifroth et al., 2013; Burlando et al., 1996; Chen et al., 2020; Li et al., 2020). Reanalysis models, however, are able to provide more than 50 years of continuous data throughout the globe (Kalnay et al., 1996; Saha et al., 2010a; Dee et al., 2011). In spite of their long time coverage, reanalysis databases present some other limitations such as a large bias (especially in mountainous areas), an overestimation of rain events with small and medium intensities, an underestimation of the most intense events, and serious difficulties to simulate small-scale physical processes (Wang et al., 2019; Bosilovich et al., 2008; Nkiaka et al., 2017). Moreover, rainfall data from reanalysis databases cannot be directly compared to point records, since model results are equivalent to spatial averages over fixed areas (pixels) (Del Jesus et al., 2015; Diez-Sierra and del Jesus, 2019).

Several authors have investigated different techniques to predict long time series of historical rainfall -covering decades in the past- to overcome the aforementioned limitations of the different sources of information. This predicted rainfall time series strengthen the robustness of technical analysis and help managers and decision makers to implement optimal strategies to problems in the fields of water resources (Li et al., 2020). This discipline is known as “Long-term Rainfall Prediction” (Gupta and Ghose, 2015) and it normally relies on regression techniques, of which a varied plethora can be found in literature, including parametric and nonparametric methods; linear and non-

* Corresponding author.
E-mail address: manuel.deljesus@unican.es (M. del Jesus).

https://doi.org/10.1016/j.jhydrol.2020.124789
Received 28 August 2019; Received in revised form 14 January 2020; Accepted 1 March 2020
Available online 05 March 2020
0022-1694/

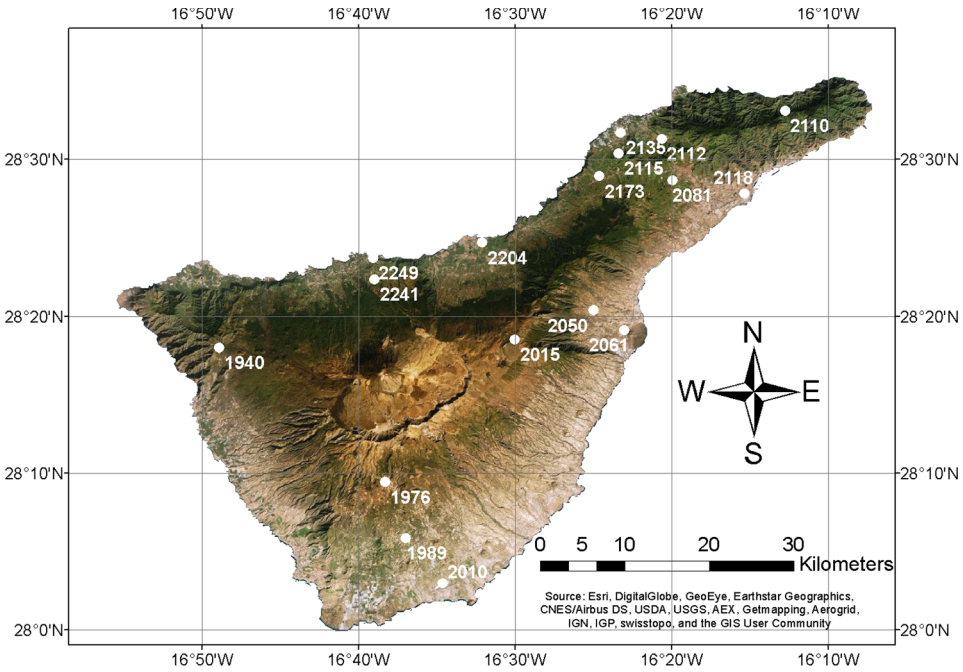

Fig. 1. Satellite view of Tenerife, location of rainfall gauges (white dots) and identification number of the gauges (four-digit white number close to the white dots). The satellite view shows *El Teide* and *Las Cañadas* in the center of the island (bright area), and the ridge that divides the island in North and South running from the Southwest to the Northeast.

linear ones (Qiu et al., 2016; Pérez-Rodríguez et al., 2012; Kannan and Ghosh, 2013; He et al., 2015; Yu et al., 2017; He et al., 2015). Generalized linear models, logistic regression and support vector machines with linear kernels are parametric models with a limited number of parameters; while k-nearest neighbors, decision trees, support vector machines with radial basis function kernels and neural networks are nonparametric models where the number of "parameters" can grow with the size of the training set (Sheskin, 2003). The increase in computational capacity has favored the use of machine learning techniques over other regression methods (Hong, 2008; Zhao et al., 2020), in part because the form of the relationship between the variables does not need to be known or assumed *a priori*. Besides, linear regressions are meant to describe linear relationships (Abbot and Marohasy, 2017) and rely on some hypothesis (normality of errors, homoscedasticity, etcetera) which may be relaxed with the use of machine learning methods.

In this paper, generalized linear models and several machine learning methods (support vector machines, k-nearest neighbors, random forests, k-means clustering and neural networks) are analyzed for long-term daily rainfall prediction, using atmospheric synoptic patterns from reanalysis databases. The main motivation of the present study is to compare the skill of the methods mentioned above to extend the temporal coverage of rainfall time series -to several decades in the past- using shorter historical records and atmospheric synoptic patterns from reanalysis database as predictors. The atmospheric patterns present a greater temporal coverage than instrumental data, allowing to effectively extend the rainfall information.

Several studies have investigated this topic in the past (Olsson et al., 2004; Sumi et al., 2012; Valverde Ramírez et al., 2005), showing that all algorithms seem to have their advantages and limitations, making the selection of the best overall algorithm difficult (Gupta and Ghose, 2015). However, none of the studies made to date performed a comparison as exhaustive as the one presented here, making use of the main methods found in literature for rainfall prediction. Besides, unlike most studies, we analyze the results not only in order to preserve some evaluation metrics, but also including several rainfall statistics such as daily variance, duration of the dry and wet spells, spatial correlation and some extremal indexes, among others, which are very important for hydrological response characterization (Serrano-Notivoli et al., 2018). Furthermore, in our study all comparisons among models are made on the basis of robust statistical tests and not merely on an analysis of differences which may not be statistically robust.

The methods presented herein can be applied worldwide (as long as instrumental precipitation data is available in the area), as reanalysis databases do normally have global coverage, and could also be extrapolated for their use in climate change studies. However, our analysis is carried out in a semi-arid climate (Tenerife, Spain), where rainfall scarcity and heterogeneity complicates the generation of accurate predictions. 17 gauges and more than 36 years of daily data are used to evaluate our methods. Predictions are done independently for each station.

In our analysis, we make use of open-license, open-source standard tools (Pedregosa et al., 2011; R Core Team, 2017) that are cross-platform, easily installable and deployable in any hardware system for any application. Some methods found in literature have been further developed for specific applications, however, they have not been considered in the analysis as their use is not widespread and are not fully cross-platform yet (Sharma et al., 2016; Adnan et al., 2019).

2. Study area and data sources

2.1. Description of the study area

Tenerife is one of the seven islands that form the Spanish archipelago of the Canary Islands. Tenerife is located in the Atlantic Ocean, 300 km west of the African coast, between parallels 28°N and 29°N, and between meridians 16°W and 17°W.

Tenerife displays a strong topographic gradient from the coast to the central part of the island. The central region of the island is a mountain range, where *El Teide* (3.718 m) is the highest peak, that separates the island in two well differentiated climatic areas: a northern region, relatively wet due to the effect of the Trade winds, and a southern region, drier due to the blocking induced by the orography (see Fig. 1). Climatic spatial heterogeneities are also present within each region, configuring a widely varied range of climates within a relatively small area (Diez-Sierra and del Jesus, 2017).

2.2. Rainfall data

Rainfall data is provided by *Consejo Insular de Agua de Tenerife* (CIATF, 2018), the water planning and managing agency for Tenerife Island. CIATF maintains a database of rainfall observations (Melián et al., 2011) that includes the observation network of *Agencia Estatal de*

Meteorología (AEMET, 2018), the Spanish national meteorological agency; and of *AgroCabildo* (AgroCabildo, 2018), a local agency for agriculture development. The CIATF database contains information for 125 gauges, combining daily and sub-daily values, with an average coverage of 15 years, although the longest series start in 1890. Only the gauges containing the best information were included in the analysis. Three criteria were considered during gauge selection: 1) gauges should cover as long a time period as possible, 2) all gauges should cover the same reference period to reduce the uncertainty associated with using different climatic time periods in the fitting procedure and 3) all climatic regions of the island should be covered by the selection, to compare the predictive skill of the models under different conditions. Following these considerations, all stations with at least 25 years of data in the period 1979–2015 were selected. Fig. 1 shows the location (white dots) and the identifier (four-digit white numbers) of the 17 rainfall gauges selected for the analysis, which are representative of the climate of the island. Rain gauges were selected to cover as much territory as possible, with the longest and most complete rainfall records for the period 1979–2015.

As we show in Table 1, precipitation in Tenerife is very scarce, with an annual average of 371 mm and 45 rainy days for the 17 gauges selected. Temporal and spatial distribution of rainfall is very heterogeneous since a very high percentage of precipitation falls in short periods of time and with a great spatial heterogeneity; largely due to the topography of the island. Seasonality is very remarkable; rainy season takes place mostly during the boreal winter (November, December and January), while the dry season covers most of the boreal summer (June, July and August).

2.3. Atmospheric data

Atmospheric variables inform statistical and machine learning methods about the state of the atmosphere; thus they constitute the main predictor for rainfall. The state of the atmosphere for Tenerife Island is properly captured (as described in previous analysis found in literature by Herrera et al. (2001) and Tullot (1959)) by the combined fields of sea level pressure (SLP), elevation of the 500 hPa geopotential surface (GH500) and elevation of the 850 hPa geopotential surface (GH850). These fields cover the area between 45° W and 5° E, and 20° N and 50° N (see Fig. 2).

Spatial atmospheric information is obtained from the global reanalysis NCEP Climate Forecast System Reanalysis (CFSR, Saha et al., 2010b) developed by the National Oceanic and Atmospheric Administration (NOAA). CFSR is a third-generation reanalysis database, generated with a global high-resolution land-ocean-atmosphere model coupled with an ice sheet model. CFSR provides 6-hourly data. The spatial resolution of the atmospheric model is 38 Km horizontally, with 64 vertical levels. The spatial resolution of the oceanic model is approximately 0.25° near the Equator, and 0.5° beyond the Tropics, with 40 vertical levels. The land model counts 4 soil levels, and the ice model 3 levels. CFSR covers the period 1979–2015.

Table 1
Gauges used in the study. The table shows the gauge identifier (ID), its elevation (Elev., m), its percentage of gaps (% gaps), its annual average rainfall ($\overline{R}_A$, mm), its percentage of dry days (p_{dry}) at a daily scale and its location within the island (North, N; South, S; Other, O). Rainfall statistics are computed for the period 1979–2015.

ID	Elev. (m)	% gaps	$\overline{R}_A$ (mm)	p_{dry}(%)	Location
1940	959.86	15.1	490.0	85.8	N
2081	633.93	0.2	545.4	73.3	N
2110	383.33	3.9	536.9	86.9	N
2112	378.33	2.6	435.0	86.2	N
2115	502.77	25.9	465.7	80.4	N
2135	122.8	25.0	323.3	84.5	N
2173	513.45	2.3	553.5	76.8	N
2204	114.06	21.5	351.9	86.6	N
2241	511.61	4.9	522.3	83.8	N
2249	49.9	1.6	315.6	85.7	N
2015	2370.16	4.7	392.9	88.3	O
2050	450.22	3.7	301.5	90.7	O
2118	47.7	0.0	230.6	84.0	O
1976	1409.1	2.5	339.0	94.6	S
1989	597.62	2.6	227.7	95.7	S
2010	71.81	4.9	129.2	93.5	S
2061	125.31	21.3	157.9	95.3	S

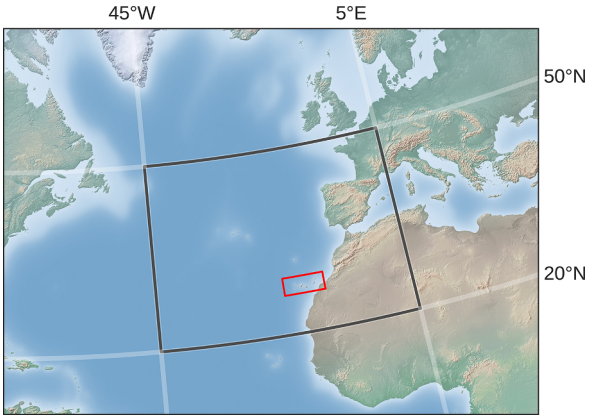

Fig. 2. Black box corresponds to the area selected for the atmospheric predictors. Red box indicates the area where Canary Islands are located.

3. Methods

Statistical and machine learning methods will be trained, or fitted, to predict daily rainfall using atmospheric predictors. Data from 17 rain gauges will be used for the analysis, keeping 80% of the data of each station for fitting the models (training), and the remaining 20% for evaluating their prediction skill (testing). 36 years long original time series for each rainfall gauge are split in two sets: the training set, containing 30 years of data; and the testing, containing the remaining 6. Models are trained independently for each station.

Machine learning techniques can be fine-tuned through hyperparameters that control their behavior. Hyperparameters may select the optimization routine used for fitting, the regularization scheme used, and the non-linear transformation applied to the input data, among other things. The training set is used to determine the optimal hyperparameters by k-fold cross-validation (Markatou et al., 2005). The optimal hyperparameters are selected from a factorial grid. The 30 years of the training set are thus further subdivided into k subsets: k-1 sets are used to train the model and the kth one is used to evaluate the performance of those hyperparameters in unseen data (validation). For each set of candidate hyperparameters, the process is repeated k times, determining an average performance of the hyperparameters and selecting the best performing set. Predictive performance is evaluated over the test set, that was not used for training, neither for selecting the optimal hyperparameters of each method (see Fig. 3). Predictive performance for rainfall occurrence and rainfall intensity is considered. Different loss functions will be used to explore their capabilities to capture different rainfall statistics.

3.1. Construction of the predictors

To construct the atmospheric fields used as predictors in this paper, CFSR data is aggregated into daily time scale (13,514 days for the whole period of analysis) over a 0.5° × 0.5° mesh (forming a 60 × 100 elements matrix). For each day, the atmospheric information consists of three (3) 60 × 100 elements matrices in total: one for SLP, one for

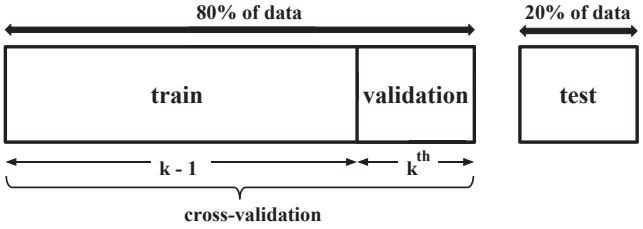

Fig. 3. Methodological scheme followed to fit and evaluate the models.

GH500 and one for GH850. Each number in the matrix represents the value of the field at a specified location; it is the component of the field on a basis vector that takes a value of one for that location, and zero everywhere else. The matrices for each day can be combined and rearranged as a vector of 1800 components (60 × 100 × 3), that are stacked together to form a matrix of 1,800 × 13,514 elements containing all the atmospheric information for the whole period of analysis.

The atmospheric information for each day could be interpreted as a data point in a 1,800-dimensional space. However, the correlations existing among the values of a variable at different locations, or among different variables at the same or at different locations, makes many of these dimensions redundant and uninformative. These correlations may disturb (and even spoil) the fitting procedure of the regression and machine learning models (Gutiérrez et al., 2004). To reduce these correlations while maintaining their discriminant power, the atmospheric patterns are standardized and then transformed through Principal Components Analysis (PCA; Abdi and Williams (2010)).

The PCA method constructs an alternative basis to represent the original data that is optimal in the sense that each basis vector captures as much variation of the original data, not captured by previous basis vectors, as possible. The first basis vector generated by the PCA method is thus aligned with the direction of maximum variation of the original data. Each successive basis vector is orthogonal to all the previous ones (it does not capture variability in the direction of previous basis vectors) and captures as much variation as possible for a single vector. The application of the PCA method renders a more workable representation of the original data, as some dimensions in the new basis may be dropped with a reduced impact in the amount of information lost. In addition, the PCA method significantly reduces the number of predictors, thus reducing also the computational cost derived from training the different statistical and machine learning models.

Many climate studies have explored different configurations of the predictors: one including direct information from the predictors from neighboring grid points, another including principal components as predictors and a last one including both (Preisendorfer, 1988; San-Martín et al., 2017). All configurations including principal components exhibit a similar overall performance, while the one that only included information of the predictor from the closest reanalysis grid box suffered from a large seasonal variability of the bias and worse predictions (Gutiérrez et al., 2013). These results indicate that the principal components from PCA constitute better predictors than the variables alone, most probably because the principal components aggregate spatial information from all the domain of the spatial predictor.

After application of the PCA method to the original matrix (1,800 × 13,514 elements), 95% of the variance of the original data set can be captured by keeping only the first 14 principal components (the components over the first 14 basis vectors generated by PCA). The atmospheric information predictor is thus a matrix with 14 × 13,514 elements; representing the time series of 13,514 days for 14 predictors (PCA components).

In addition to the atmospheric variables we incorporate seasonality as a predictor. Seasonality introduces a non-stationarity in the distribution of rainfall that must be considered for accurate predictions. A simple way to include this non-stationarity is to use as predictor some function that mimics the shape of the average monthly rainfall throughout the year. The simplest analytical function able to reproduce that shape is a combination of a sine ($y_1 = \sin(2\pi t)$) and a cosine ($y_2 = \cos(2\pi t)$) of period (t) 12 months. The sine and cosine functions (y_1 and y_2) will be two additional predictors given to the statistical and machine learning techniques, informing these methods about the differences existing among months (Méndez et al., 2007).

3.2. Statistical and machine learning methods

Statistical methods traditionally used to predict precipitation are based on classical Linear Regression Models (LRM), and Generalized Linear Models (GLM) that allow for response variables presenting non-Gaussian error distributions (Coe and Stern, 1982; Stern and Coe, 1984). However, with increasing computational power, machine learning approaches have extended the options of rainfall prediction techniques (Qiu et al., 2016). Machine learning techniques have two major advantages: they do not normally require any assumption about the distribution of errors and the form of the relationships between predictors and predictands does not need to be known *a priori* (Appelhans et al., 2015). The intermittency of rainfall at daily time scales (the succession of wet and dry intervals of finite duration) makes it more convenient to split the prediction problem into two steps. Thus, rainfall occurrence and rainfall intensity are modeled separately (Olsson et al., 2004). For this reason, statistical and machine learning techniques are separated into two categories: *classifiers* that deal with categorical data (Rain and No Rain), aimed at predicting rainfall occurrence; and *regressors* that deal with continuous data, aimed at predicting rainfall intensity.

Table 2 enumerates all the statistical and machine learning models used in this analysis. The first column shows the model name, the second column corresponds to the abbreviation (Abbr.), the third column shows whether the method is used for classification (C), regression (R) or both (C/R), the forth column enumerates the hyperparameters optimized during the analysis, and finally, the fifth column lists the values of the hyperparameters explored in the model optimization step. The implementation of the methods found in the *stats* library (R Core Team, 2017) and the *Scikit-learn* library (Pedregosa et al., 2011) were used in this work.

In the present work two statistical (GLMs) methods are used. On the one hand, GLM-L, which corresponds to Classical Lineal Regression but assuming a logarithmic relation between predictor and predictand. The logarithm link function transforms the data to increase their similarity to a Gaussian distribution, improving the quality of the fit. On the other hand, GLM-G, which assumes that errors follow a gamma distribution. The gamma distribution is commonly used to model rainfall because the distributions of precipitation amount tend to be strongly skewed at daily time scales (Ben Alaya et al., 2017; Stephenson et al., 1999; Yang et al., 2005).

Six different machine learning methods are used in the analysis. Random Forest (RF) is an ensemble learning method that constructs a multitude of decision trees at training time and predicts the mode of the predicted classes (classification) or the mean of the predictions (regression) of the individual trees. Two hyperparameters are analyzed for optimization: *N_estimators*, that corresponds to the total number of trees in the forest, and *min_samples_leaf*, that is the minimum number of samples required to be at a leaf node, a node over which the mode or the mean are computed. K-Nearest Neighbors (k-NN) is a non-linear method whose predictions are computed through the weighted mode (classification) or the weighted mean (regression) of the k-nearest points to the one being predicted. Two hyperparameters are analyzed for optimization: *n_neighbors*, that corresponds to the number of neighbors used in the prediction, and *weights*, that is the weight function used in prediction. *Uniform* means that all points in the neighborhood are equally weighted while *distance* means that points are weighted by the inverse of their distance. Neural Networks (NN) learns

Table 2
Statistical and machine learning methods used in the study. Method names appear in the first column; the second column corresponds to abbreviations (Abbr.); whether the method is used as classifier (C), regressor (R) or both (C/R) is shown in the third column; the fourth column specifies the hyperparameters tested during grid search model selection and the fifth column lists the values of the hyperparameters explored in the model optimization step.

Model	Abbr.	C/R	Grid Search	
			Parameters	Values
Generalized Linear Models (family=gaussian, link=log)	GLM-L	R	None	None
Generalized Linear Models (family=gamma, link=log)	GLM-G	R	None	None
Logistic Regression	LR	C	solver	sag,newton-cg, lbfgs,liblinear
			penalty	l1, l2
			C	0.1,1,5,10,25,50,75,100,200,1000
Random Forest	RF	C and R	n_estimators	1,2,5,10,20,40,60,80,100,200
			min_samples_leaf	1,2,4,6,10,15,30,50
k-Nearest Neighbours	k-NN	C and R	n_neighbors	1,2,3,4,5,6,7,8,9,10,15,20,25,30,35,40,45,50
			weights	uniform, distance
Support Vector Machines	SVM	C	C	0.1,1,2,5
			kernel	rbf
			γ	0.005, 0.01, 0.025, 0.05
		R	C	25,50,75,100,150,200,250,300
			kernel	rfb
			ϵ	0.001,0.1,0.2,0.5,1,2,5,10,20
			γ	0.01,0.025,0.05,0.075,0.1
Weather Typing (K-means)	WT	C and R	n_clusters	4,9,16,25,36,49,64,81,100,121,144
Neural Networks	NN	C	solver	lbfgs, sgd, adam
			alpha	0.0001,0.001,0.1,10,1000
			hidden_layer_sizes	N:2,3,4,5,6,7,10; L:1, 2, 3
			activation	identity, logistic, tanh, relu
			max_iter	2000
		R	solver	lbfgs, sgd, adam
			alpha	0.0001,0.001,0.1,10,1000
			hidden_layer_sizes	N: 2,3,4,5,6,7,8,9,10,15,20; L:1, 2, 3
			activation	identity, logistic, tanh, relu

a mapping between a series of input features (input layer) and an output (output layer) through linear combinations and non-linear transformations of the inputs received by every neuron in the network (hidden layer). Three hyperparameters are analyzed for optimization: *Hidden_layer_sizes*, that is the number of neurons and layer used, *activation* that defines the non-linear function that each neuron uses to transform the linear combination of inputs it receives, and *alpha*, a term controlling regularization strength by penalizing weights with large magnitudes. Support Vector Machines (SVM) perform classification and regression tasks by finding the hyperplanes that maximize the margins, the minimum gap separating samples belonging to different groups. The closest values to the classification margin are known as support vectors. Radial Basis Function (RBF) kernels are used to ensure separation is possible in non-linear spaces. Three hyperparameters are analyzed for optimization: γ, that is the radius of the area of influence of the support vectors, C, that can be seen as the inverse of the regularization strength and ε which defines a margin of tolerance where no penalty is given to errors. Logistic Regression (LR) classifies binary outcomes modeling the logit transformed probability of occurrence through linear regression. Three hyperparameters are analyzed for optimization: *Solver*, that is the algorithm used in the optimization problem, *penalty*, that is the regulation function applied to the weights of the regression, and C, that can be seen as the inverse of the regularization strength (like in SVM). Finally, Weather Typing (WT), which divides the input space into regions, and uses the mode or mean of the region in which the candidate point lies to solve the classification or regression problem. In this specific application, weather types are computed using the k-means algorithm (Camus et al., 2011), turning weather types into representative synoptic atmospheric patterns (Diez-Sierra and del Jesus, 2017). WT is the only method in which predictions are done simultaneously for all the gauges.

A more detailed description of the hyperparameter selection as well as the sensitivity of the models to each of them can be found at Scikit-learn (2019).

3.3. Loss functions

A loss function measures the distance between the value estimated by a statistical or machine learning method and the objective value the model is supposed to predict. The fitting procedure of any statistical or machine learning method involves finding the values of the parameters that optimize the value of the loss function. In most cases, the loss function used to fit regression models is the sum of squared errors, either with a focus on bias or variance. An advantage of machine learning methods is that the loss function can be easily customized so that values predicted by the model may minimize any convenient function.

In the present work, two different loss functions are used depending on the objective of the fitting procedure. Classifiers, used mainly to predict rainfall occurrence (or probability), are fitted using f-score:

$$\text{f-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (1)$$

with precision being defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

and recall as

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

where TP means true positives, the number of observed rainy days correctly predicted as rainy days by the model; FP means false positives, the number of observed non-rainy days predicted as rainy days by the model; and FN means false negatives, the number of observed rainy days predicted as non-rainy days by the model.

Precision measures the probability that a predicted rainy day by the model corresponds with a real rainy day (a true positive), while recall measures the probability that a real rainy day is correctly predicted by the model (a true positive). A value of one (1) in both measures represents a perfect score, while zero (0) is the minimum value for both metrics. A model that only predicts rainy days when rainfall really occurs, but misses some observed rainy days, predicting a non-rainy day, would have a perfect precision (a value of 1), but would have a recall smaller than one (1) as some events would have not been properly recognized (false negatives exist). A model that predicts rainy days when rainy days really occur but also in some non-rainy days, would have a perfect recall, but a precision smaller than one (1), as false positives are present in the prediction.

The f-score metric is very useful to compare the skill of model predictions when working with imbalanced datasets (datasets in which one class is much more frequent than the others). This is the case of rainfall in Tenerife, where rainy days are rare, with an average value for the 17 gauges of 15% of the days. However, the proportion of rainy days can vary from 25%, for some gauges located in the north of the island, to 5% for other gauges located in the south.

Regressors, which aim to predict the intensity of rain, are evaluated using the Root Mean Squared Error (RMSE) as loss function:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} (x_{1,t} - x_{2,t})^2}{n}} \quad (4)$$

where x_1 and x_2 correspond to the observed and the simulated rainfall series, respectively.

Although not used as a loss function, bias (BIAS) and Pearson correlation (R), as defined in Eqs. (5) and (6), respectively, are also used as evaluation metrics for the models.

$$BIAS = \frac{\sum_{t=1}^{n} (x_{1,t} - x_{2,t})}{n} \quad (5)$$

$$R = \frac{\sum_{t=1}^{n} x_{1,t}x_{2,t} - n\bar{x}_1\bar{x}_2}{\sqrt{\sum_{t=1}^{n} x_{1,t}^2 - n\bar{x}_1^2}\sqrt{\sum_{t=1}^{n} x_{2,t}^2 - n\bar{x}_2^2}} \quad (6)$$

In Eqs. (5) and (6), $x_{1,t}$ and $x_{2,t}$ correspond to the observed and the simulated rainfall series, respectively, and n is the sample size.

3.4. Grid search

Before evaluating the predictive performance on the original test set, the optimal hyperparameters for each class of models shown in Table 2 are determined. Hyperparameters are chosen from a factorial grid (the fourth column of Table 2 shows the hyperparameters tuned for each model and the fifth column shows the explored values). The specified model is trained with different hyperparameters and evaluated over the training set using k-fold cross validation (Markatou et al., 2005). The k-fold cross-validation approach avoids overfitting and the appearance of spurious effects of any particular partition of the input data. k-fold cross validation implies dividing the training set in k equal-sized subsets, using k-1 subsets to train the model, and evaluating the performance on the kth subset (the one not used for model training). The procedure is repeated k times, using each of the k subsets one time to evaluate performance. The average performance of the k trainings is assigned to the hyperparameters, and the optimal hyperparameters are the ones that maximize the performance of the model. In this study, k is taken equal to 5, so that 80% of the original training data are used for training and the remaining 20% used for validating the hyperparameters. Each training set is thus split in five subsets; using 4 sets for fitting the model and 1 for validation.

Grid search involves, in the present study, fitting over 25,000 different models. The number 25,000 results from the combination of different models (regressors and classifiers) with their respective hyperparameters and the 17 gauges.

3.5. Model evaluation

As we mentioned in the subSection 3.2, rainfall occurrences and rainfall intensity are modeled separately. On the one hand, the models defined as classifiers in Table 2 are used to reconstruct the entire time series of rainfall occurrence for the 17 gauges independently. To this end, the continuous time series of precipitation are transformed into binary series (0 = NoRain, 1 = Rain) before training. F-score is the loss function applied for all classifiers (Eq. (1)). On the other hand, the models defined as regressors in Table 2 are used to reconstruct the entire time series of rainfall for the 17 gauges independently. All values larger than 0.1 mm/day are used in the analysis. The loss function in this case is RMSE (see Eq. (4)). Seasonality, sine and cosine functions, is included as a predictor for the rainfall occurrence and intensity.

4. Results

4.1. Hyperparameter sensitivity analysis

During the cross-validation procedure, several sets of hyperparameters are tested to find the optimal one. The performance of all the tested hyperparameters sets, both for classifiers and regressors, are shown in Fig. 4. The left-hand panel of the figure shows the classifier's performance, through f-score; while the right-hand panel shows the regressor's performance, through the Pearson correlation coefficient (R). The figure shows the distribution of f-score and R values for the different models and for every rainfall station by means of histograms and kernel density estimates of the probability density function. Both f-score and R have a range of values from 0 to 1. The vertical dashed line represents the middle of the range; the value of 0.5.

An analysis of the distributions for the classifiers reveals that the performance of logistic regression (LR) and support vector machines (SVM) does not depend much on the value of the hyperparameters used. Indeed, the distributions of f-score for these two models are almost Dirac's deltas. These models will therefore present a skill mostly dependent on the characteristics of the covariates and the predicted series, but cannot be easily adapted by tuning their hyperparameters. A similar case, although not so extreme, would be weather typing (WT), that shows little sensitivity to hyperparameter selection. However, its performance tends to be lower than the two previous methods.

On the opposite side of the spectrum, neural networks (NN) presents a widely spread distribution, indicating a high sensitivity to hyperparameter tuning. NN performance is highly dependent on the hyperparameter values used during training. Indeed, their performance may be the worst of all models, as can be seen by the amount of

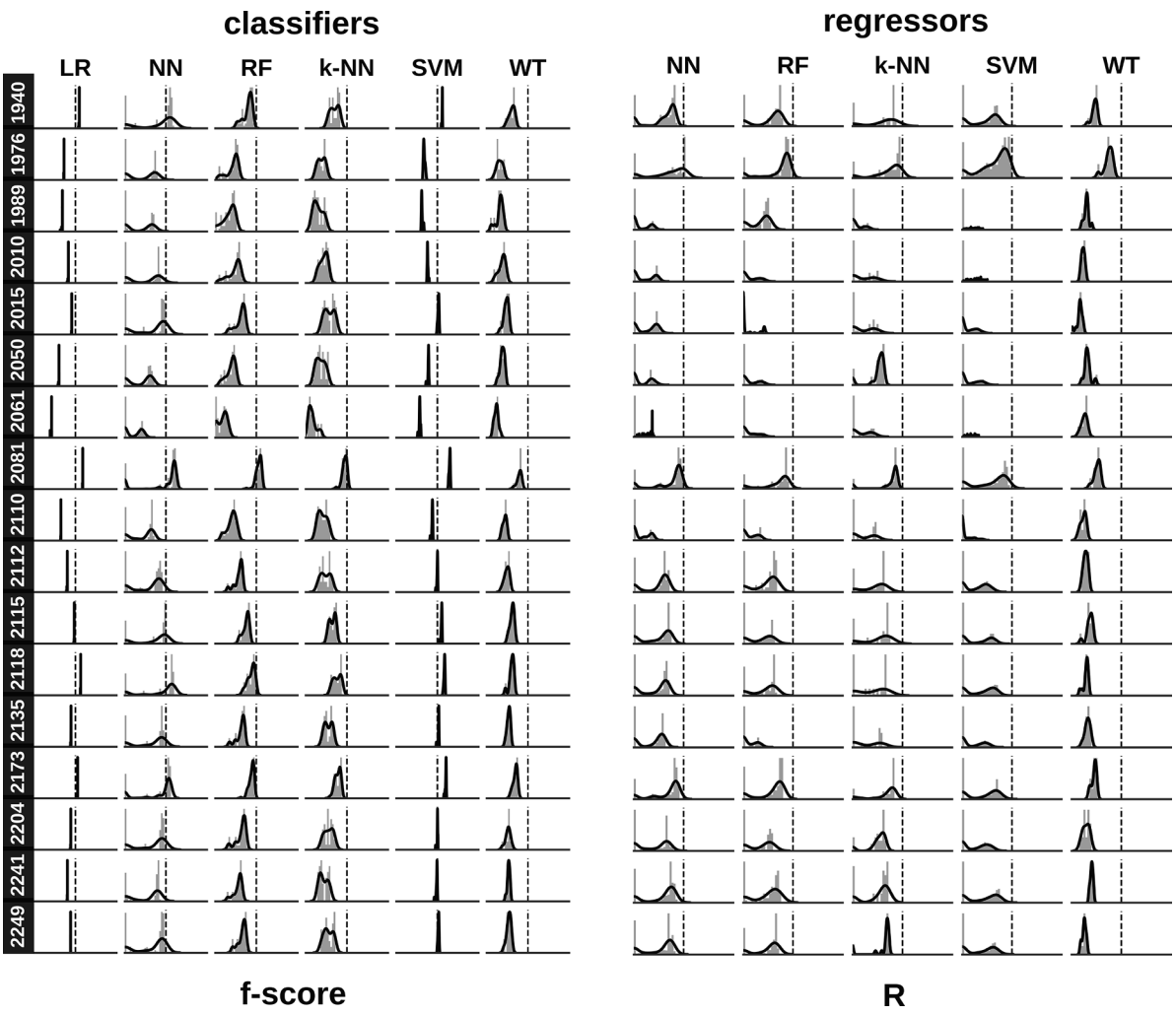

Fig. 4. Performance skill for the whole combination of hyperparameters presented in Table 2 for the validation set. Columns and rows represent models and gauges, respectively. Histograms and kernel distributions are used to represent the results of f-score (left panel) and R (right panel) for the classifier and regressors, respectively. X-axis goes from 0 to 1. Vertical dashed line corresponds to the value of 0.5.

probability concentrated around zero skill, but with a proper hyperparameter calibration, they can provide the best predictions over all models.

Finally, random forests (RF) and k nearest neighbors (k-NN) constitute a middle ground among models. Their performance is sensitive to the values of the hyperparameters, but their spread is narrower than for NN.

The right-hand side of Fig. 4 shows that, in general, regressors are more sensitive to hyperparameters values than classifiers. Except for WT, that shows a smaller variation on performance as hyperparameters are varied, all the other models present wide distributions, indicating a high sensitivity to hyperparameter tuning.

Another important result of this analysis is that the optimal hyperparameter set for a given method changes for each gauge (see Tables A.11 and A.12 in Supplementary materials); the only exception being WT, for which the same number of clusters maximized the Dunn index criterion (Baarsch and Celebi, 2012) for all gauges. Despite this variability, some conclusions can be derived from the analysis.

Regression requires more complicated models than classification, as it could be expected. For instance, RF requires a higher number of trees (*N_estimators*) and a higher minimum number of samples in each leaf node (*min_samples_leaf*) to predict rainfall intensity than rainfall occurrence. A similar result is observed with the number of neighbors required in the prediction (*n_neighbors*) for k-NN; a higher number of neighbors is required to predict rainfall intensity than rainfall occurrence. In the case of SVM, two properties are pursued: a hyperplane with the largest minimum margin (C), and a hyperplane that correctly separates as many instances as possible. For large values of the regularization term C, the optimization will choose a smaller-margin hyperplane, while a lower C will encourage a larger margin, and a simpler decision function. Cherkassky and Ma (2004) suggests that with an optimal choice of the hyperparameter ε, which defines the margin of tolerance where no penalty is given to errors, the value of C has negligible effects on the performance (as long as C is larger than a certain threshold determined from the training data). Thus, we recommend to first define the threshold of C, and then set the value of hyperparameter ε. SVM, in *Scikit-learn*, presents a default value of ε equal to 0.1. However, the magnitude of ε must be increased to reduce BIAS. Larger values of ε are normally required especially for very large and/or noisy training sets (Mattera and Haykin, 1999). NN uses parameter α for regularization, which helps in avoiding over-fitting by penalizing weights with large magnitudes.

As in SVM, NN, in *Scikit-learn*, presents a default value of α equal to 0.0001. We found that the hyperparameter α must be increased several orders of magnitude to avoid over-fitting in some of the gauges. Hyperparameter *Hidden_layer_sizes* in NN defines the number of neurons and layers used. The values in the parenthesis correspond to the number of neurons and layers, respectively (see Tables A.11 and A.12) As we can see, the number of layers required are 1 for all the gauges, except in 2 of the cases. In contrast, the number of neurons can vary from 2 to 20.

Table 3
Average f-score for the 17 gauges. Training and testing performances are showed.

f-score											
LR		NN		RF		k-NN		SVM		WT	
train	test	train	test	train	test	train	test	train	test	train	test
0.38	0.37	0.42	0.39	0.63	0.33	0.8	0.32	0.40	0.36	0.27	0.26

Table 4
Average R for the 17 gauges. Training and testing performances are showed.

R													
GLM-L		GLM-G		NN		RF		k-NN		SVM		WT	
train	test	train	test	train	test	train	test	train	test	train	test	train	test
0.40	0.29	0.37	0.32	0.41	0.37	0.60	0.33	0.93	0.34	0.39	0.35	0.18	0.18

Table 5
Results of t-tests for the model skill predicting rainfall occurrence measured over the f-score metric. An S indicates that there is a significant difference between the results of both models, while an N indicates that there is not.

	K-NN	LR	NN	RF	SVM
LR	S	–	–	–	–
NN	S	S	–	–	–
RF	N	S	S	–	–
SVM	S	N	S	S	–
WT	S	S	S	S	S

Table 6
Average recall metric for the 17 gauges and the thresholds of 0.1, 5, 20 and 40 mm/day.

	Thresholds (mm/day)				
	0.1	5	10	20	40
LR	0.37	0.47	0.51	0.56	0.59
NN	0.39	0.51	0.54	0.57	0.63
RF	0.34	0.41	0.45	0.48	0.51
k-NN	0.30	0.39	0.43	0.46	0.46
SVM	0.36	0.45	0.49	0.50	0.53
WT	0.26	0.31	0.34	0.34	0.36

Activation functions in NN introduce non-linear properties to our model. Rectified Linear units, *Relu* ($f(x) = max(0, x)$), and hyperbolic tangent, *tahn* ($f(x) = tanh(x)$), activation functions are selected in most of the gauges. *Identity* activation function, that establishes linear relationships, is only selected twice.

Tables 3 and 4 show the average skill during training and testing for all methods. Results are calculated as the average result of the 17 gauges. As expected, most models present better skill in training that in testing. As mentioned in Section 3, in order to avoid over-fitting we apply cross-validation on 80% of the original training set; cross-validation serves to select the best hyperparameters. Then we check that the skill of the methods remains similar in the remaining 20% of unseen data (test). Despite this decomposition of the original data, RF and k-NN over-fit, as seen by the difference of the skill between training and testing. In the case of k-NN, this fact can be explained because the optimal number of neighbors selected for prediction is equal to 1 for several gauges. For these gauges, the k-NN model selects, from the training set, the rainfall value associated to the most similar atmospheric synoptic pattern for each prediction; reaching a perfect fit during training. Something similar happens with RF and the hyperparameter *min_samples_leaf*. When the value selected in the cross-validation is too small, RF predicts values very similar to those seen during training. To avoid over-fitting, the hyperparameters *N_estimators* and *min_samples_leaf*, in the case of RF, and *n_neighbors* for k-NN, should be forced to take higher values, even when the training skill is degraded. However, attention must be paid in order not to force the model to under-fit. The high variability showed in Fig. 4 demonstrates that it is essential to calibrate the hyperparameters prior to any application.

4.2. Modeling rainfall occurrence

The average performance of every model for predicting rainfall occurrence is shown in Table 3. NN, with an average f-score close to 0.4, appears as the best performing model, closely followed by LR and SVM. The complete set of f-scores for every gauge can be found in Table A13 in Supplementary materials. RF and k-NN provide worse results whereas the WT based method shows the worst predictive performance, far below the rest of the models. The skill of the prediction is substantially better for gauges located in the North of Tenerife that in the South (see Table 1). This is probably because classifiers are more able to capture the underlying relationship in frontal precipitations, which dominates in the North of the island, than the convective processes prevailing in the South. In general, the performance of NN classifiers seems superior on all gauges, except for a few cases, like gauge ID:2115 where SVM did a better job than NN.

In order to verify the hypothesis of the superior skill of NN classifiers, a test of significance was carried out. Multiple t-tests were carried out to verify which models presented significant differences among them, making use of the Holm correction method (Holm, 1979) in order to counteract the problem of multiplicity. T-tests were carried out over the f-score metric. Table 5 shows the results of these tests, indicating where the differences in these two-model comparisons were significant. It can be seen that NN presents significantly better results than the other models, allowing us to conclude that NN are the best performing model for predicting rainfall occurrence. LR and SVM are placed second, presenting significantly better results than RF, k-NN and WT.

Another important metric to evaluate is the performance of a model predicting exceedances over a threshold, as this will indicate the skill of the model predicting the most intense rainfall events. Rainfall occurrence over four thresholds 0.1, 5, 20 and 40 mm/day is analyzed. Recall measures how many of the positive samples (observed rainy days with

Table 7
Observed (Obs) and simulated transition probabilities (ϕ_{DD} and ϕ_{WW}). Table shows the mean (μ), median (M) and the 25th (Q_1) and 75th (Q_3) percentiles calculated from the 17 gauges.

	ϕ_{DD}						ϕ_{WW}					
	Obs	LR	NN	RF	k-NN	WT	Obs	LR	NN	RF	k-NN	WT
μ	86.6	86.6	86.5	85.9	87.9	87.0	43.8	33.8	36.1	29.6	29.6	23.7
M	86.2	86.1	86.1	85.5	87.6	86.4	42.1	34.7	36.3	30.2	29.2	25.4
Q_1	84.0	84.0	83.8	83.2	85.7	84.8	39.3	29.3	30.9	26.7	24.2	19.1
Q_3	90.7	90.6	90.7	90.1	92.1	90.9	51.4	39.0	39.3	33.1	33.5	27.4

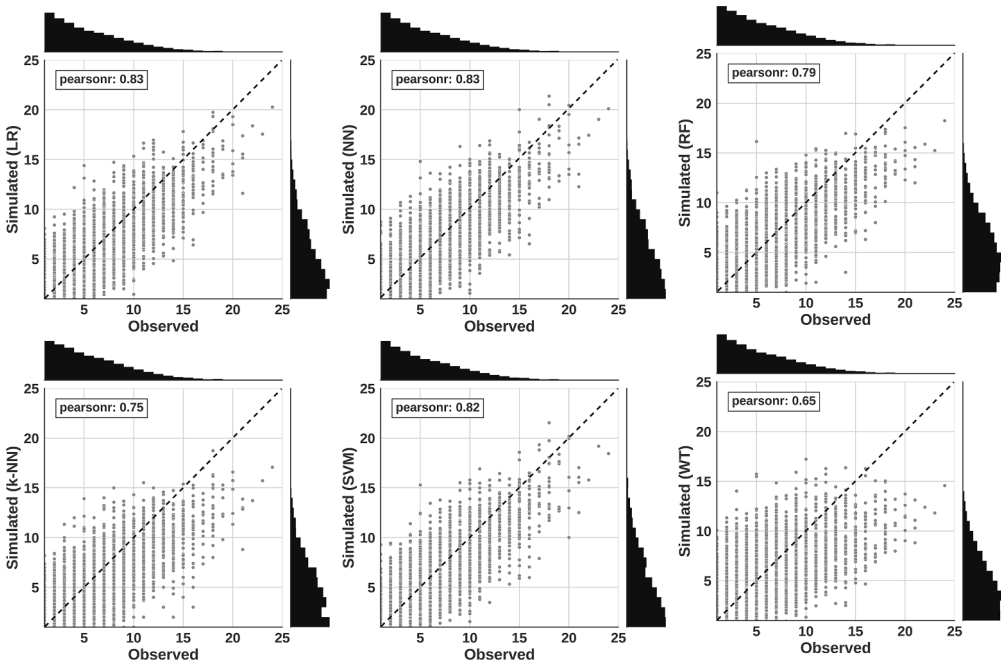

Fig. 5. Observed and simulated number of wet days (prcp ⩾ 0.1 mm) per month represented by joint probability distributions. 17 gauges are shown together. Dashed line indicates a perfect fit.

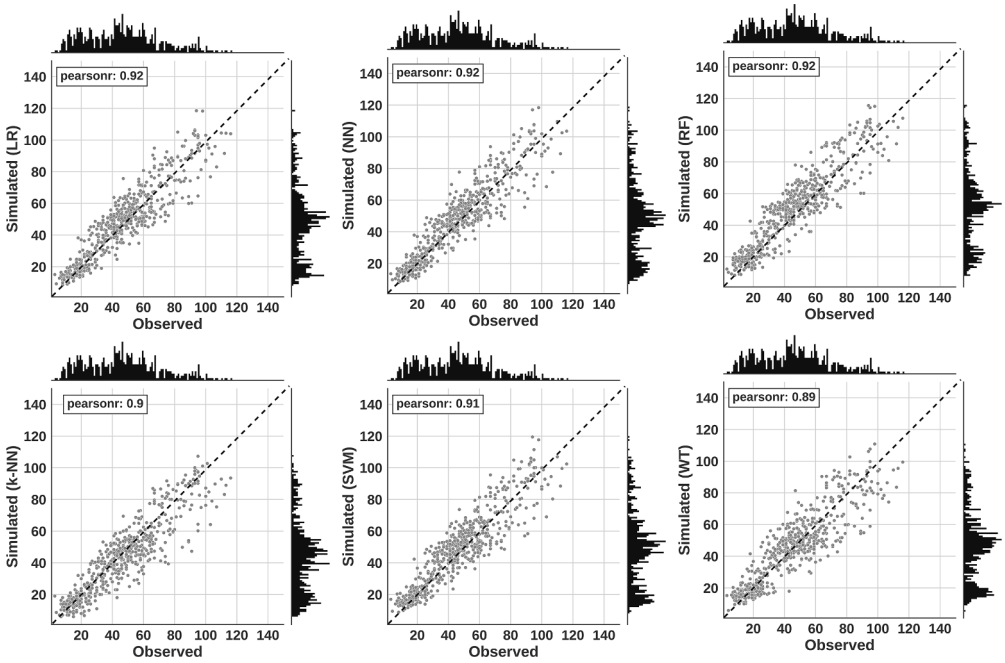

Fig. 6. Observed and simulated number of wet days (prcp ⩾ 0.1 mm) per year represented by joint probability distributions. 17 gauges are shown together. Dashed line indicates a perfect fit.

Table 8
R and RMSE metrics. Table shows the mean (μ), median (M) and the 25th (Q_1) and 75th (Q_3) calculated from the 17 gauges.

	GLM-L		GLM-G		NN		RF		k-NN		SVM		WT	
	R	RMSE	R	RMSE	R	RMSE	R	RMSE	R	RMSE	R	RMSE	R	RMSE
μ	0.29	13.6	0.32	13.2	0.37	12.6	0.33	12.8	0.34	12.8	0.35	12.7	0.18	13.7
M	0.31	11.7	0.33	11.6	0.36	10.7	0.33	10.9	0.32	10.8	0.33	10.8	0.16	11.5
Q_1	0.2	9.5	0.26	9.9	0.31	9.1	0.27	9.4	0.28	9.3	0.29	9.2	0.14	10.0
Q_3	0.4	17.8	0.38	15.7	0.44	15.2	0.38	14.9	0.38	15.2	0.42	15.2	0.23	16.9

Table 9
Results of the t-tests for the model skill predicting average rainfall intensity measured over the RMSE. An S indicates that there is a significant difference between the results of both models, while an N indicates that there is not.

	GLM-G	GLM-L	k-NN	NN	RF	SVM
GLM-L	N	–	–	–	–	–
kNN	N	N	–	–	–	–
NN	S	S	S	–	–	–
RF	N	N	N	S	–	–
SVM	S	S	N	S	S	–
WT	S	S	S	S	S	S

an intensity larger than a specific thresholds) are captured by the positive predictions (simulated rainy day). Recall has the advantage of not involving the false positives (type I errors) values in the analysis. Table 6 shows the average recall metric for the 17 gauges and the selected thresholds. The table shows that the recall metric is higher, for all classifiers, as the value of the threshold increases, reaching 0.63 for the NN classifier for the threshold of 40 mm/day. The increase of the recall metric with the increasing thresholds indicates that larger rainfall events are easier to predict, as they are more dependent on atmospheric synoptic situations.

Rainfall persistence is another important characteristic linked to rainfall occurrence. It can measured by transition probabilities, that indicate the probability of having a dry day followed by another dry day (ϕ_{DD}), and having a wet day followed by another wet day (ϕ_{WW}). Table 7 shows the mean (μ), median (M) and the 25th (Q_1) and 75th (Q_3) percentiles calculated from the 17 gauges. The statistics give us an idea of the distribution of values for the 17 stations. Although the mean (μ) and median (M) might seem redundant, we decided to show both statistics because the median (M) is more robust in the presence of outliers than the mean. As we can see in Table 7, all methods are able to preserve the observed (Obs) ϕ_{DD} with very small errors. However, they underestimate the observed values of ϕ_{WW}. If fact, some of them give large errors when comparing the predicted and the observed values of ϕ_{WW}, such as WT, k-NN and RF methods. NN, LR and SVM, again, present the best average results, however they still underestimate ϕ_{WW} in approximately 25% of the observed values. NN presents the smallest error predicting ϕ_{WW} mainly due to its better ability to predict rainy days. As Table 7 demonstrates, all the statistics (μ, M, Q_1 and Q_3) calculated from the simulated series underestimate the observed ones, which implies that the simulated distribution of ϕ_{WW} is displaced to lower values. The complete set of observed and simulated transition probabilities (ϕ_{DD} and ϕ_{WW}) for every gauge can be found in Table A.14 in Supplementary materials.

Interesting results are found when the skill of classifiers to reproduce the number of wet days per month and year is analyzed. Fig. 5 shows a scatter plot of the observed and simulated number of wet days per month for all classifiers and gauges. As we can observe, NN, LR and SVM present, on average for all the gauges, the best skill with values of R close to 0.83. RF and k-NN show slightly worse results, while WT presents the worst skill with a value of R equal to 0.65. NN, LR and SVM are the only methods able to simulate months with more than 20 wet days, which is exceptionally unusual in Tenerife, although possible. Besides, they show less biased distributions with more similar shapes to the observed ones. Similarly, Fig. 6 shows a scatter plot of the observed and simulated number of wet days per year for all classifiers and gauges. As we can observe, the skill of the models improves when the results are evaluated at annual temporal aggregation. All models except WT present values of R higher or very close to 0.9. All models, also, are able to predict properly the extremes of the distribution which correspond to years with more than 100 rainy days and less than 20, respectively. The results shown in Figs. 5 and 6 indicate that, even when the daily time series of rainfall occurrence is not perfectly predicted using machine learning methods and synoptic patterns, the prediction is acceptable, and even good, when aggregated over larger time scales. The prediction is consistent in statistical terms, and therefore is more useful for those applications where the prediction of the exact timing of rainfall is less relevant.

4.3. Modeling rainfall intensity

Table 8 shows the result of the Pearson correlation coefficient (R) and root mean square error (RMSE) for all regressors; both R and RMSE were computed on the subset of wet days (prcp $>$ 0.1 mm) in order to reduce the effects of the varying occurrence frequencies. The statistics mean (μ), median (M) and the 25th (Q_1) and 75th (Q_3) percentiles calculated from the 17 gauges, are summarized in the table (the results for the 17 gauges are shown in Table A.15, in Supplementary materials). NN shows the highest average value of R (0.37) and the smallest average value of RMSE (12.6 mm/day). SVM, k-NN and RF give slightly worse values of R and RMSE than NN, while WT presents the worst result. Statistical methods (GLM-L and GLM-G) perform, in general, worse than machine learning models, with the exception of the WT approach. GLM-L and GLM-G show higher values of R for the median (M) than for the mean (μ), which means that in those stations with lower values of R the statistical methods present less skill. If we observe the results for the percentiles (Q_1 and Q_3), we can see that NN presents again the best results of all the regressors. R exceeds the value of 0.44 for 25% of the gauges.

When we analyze the results for the 17 stations separately (see Table A.15 in Supplementary materials), we can see that as in the case of the rainfall occurrence, and pretty much for the same reasons, most of the gauges with the best results (higher values of R and lower RMSE) are located in the North of the island (ID: 2241, 2173, 2081, 2204, 1940). It is interesting to note that some stations separated by a few kilometers present very different values of R such as the stations 2173 and 2135. These differences, however, might be explained by local conditions, specifically by the difference in elevation between them (around 400 m), which results in separate rainfall regimes.

The hypothesis of the superior skill of NN predicting average rainfall intensity is tested using a multivariate ANOVA (MANOVA) test (Muller and Peterson, 1984) for paired samples, under the null hypothesis that all models provided equally good predictions, when measured through

Table 10
Observed and simulated values o for the following statistics: daily mean (μ), daily variance (σ_2), R20 (number of days with precipitation over 20 mm) and RX1 (maximum 1-day precipitation).

	μ								σ^2								$RX20$								$RX1$							
	Obs	GLM-L	GLM-G	NN	RF	k-NN	SVM	WT	Obs	GLM-L	GLM-G	NN	RF	k-NN	SVM	WT	Obs	GLM-L	GLM-G	NN	RF	k-NN	SVM	WT	Obs	GLM-L	GLM-G	NN	RF	k-NN	SVM	WT
1940	9.6	9.7	9.9	9.8	9.6	8.4	8.3	9.6	225	48.1	81.3	65.3	34.3	23.9	36.5	45.1	220	536	137	184	98.0	59.0	80.0	105	130	85.5	97.8	52.5	32.6	29.6	36.2	50.5
1976	17.1	17.0	18.6	16.9	17.0	16.0	19.1	17.6	685	315	556	197	154	99.0	167	245	169	395	193	245	236	203	302	251	182	166.5	240.0	68.3	65.3	58.4	72.8	78.5
1989	14.7	14.6	14.9	16.1	14.7	14.9	14.4	15.4	324.1	80.4	83.7	64.5	35.8	21.2	20.0	85.2	126	388	106	167	101	80.0	67.0	175	130	100	75.9	46.4	37.4	30.7	34.9	80.0
2010	5.9	4.6	6.1	5.9	6.0	6.2	6.0	6.1	126	83.8	24.6	17.3	9.3	9.3	4.3	17.7	44.0	84.0	16.0	0.0	0.0	1.0	0.0	2.0	136	150	54.6	15.8	15.1	20.3	13.0	20.0
2015	9.5	6.7	9.5	9.4	9.7	8.0	7.6	9.1	388	120	76.6	43.5	30.9	14.3	24.1	31.4	168	280	97.0	97.0	89.0	21.0	32.0	56.0	337	199	97.7	51.6	34.0	28.9	39.3	35.1
2050	9.0	8.1	9.1	9.0	8.9	7.6	8.3	9.0	307	61.7	76.1	26.9	25.8	24.3	24.8	38.5	148	300	85.0	51.0	44.0	35.0	40.0	51.0	194	69.9	85.3	21.4	28.2	35.3	34.3	40.9
2061	9.2	8.6	9.5	9.3	9.3	8.4	9.6	9.2	240	134	50.0	12.3	23.5	9.6	15.4	86.1	54.0	109	32.0	5.0	33.0	4.0	10.0	49.0	130	174	66.1	23.5	22.4	22.8	33.3	71.9
2081	6.0	6.1	6.2	6.0	6.1	5.3	6.3	6.0	108	23.3	47.1	30.3	19.6	17.0	17.7	14.9	228	427	127	98.0	57.0	27.0	45.0	6.0	134	57.1	103	33.8	28.2	30.7	30.9	33.2
2110	11.2	11.0	11.3	11.2	11.2	10.3	10.0	11.2	236	40.2	35.6	20.9	21.5	15.7	17.8	27.4	203	772	121	81.0	103	38.0	45.0	125	216	85.3	63.1	36.9	30.2	29.0	33.0	54.0
2112	8.6	8.6	8.6	8.6	8.7	8.3	8.7	8.6	112	15.7	17.9	16.5	14.0	13.0	11.7	13.1	164	478	47.0	0.0	18.0	19.0	9.0	5.0	145	35.3	44.6	19.0	23.4	24.6	27.8	23.2
2115	6.7	6.7	6.8	6.6	6.7	6.2	6.9	6.5	98.7	18.9	28.9	15.1	15.8	10.3	7.6	12.7	133	277	63.0	0.0	20.0	2.0	1.0	2.0	97.0	39.3	56.1	18.3	24.9	21.7	22.5	32.0
2118	4.5	4.5	5.0	4.5	4.5	3.9	3.9	4.5	112	13.0	56.0	17.3	11.7	7.1	12.5	10.7	87.0	123	58.0	0.0	1.0	0.0	4.0	6.0	233	40.0	126	17.8	21.0	16.3	22.1	22.7
2135	5.8	5.8	5.9	5.7	6.0	5.5	5.4	5.6	80.0	13.9	15.8	9.1	10.9	6.5	9.1	9.6	72.0	159	27.0	0.0	5.0	0.0	0.0	1.0	127	30.7	35.9	19.4	25.1	18.6	19.9	20.2
2173	6.6	6.6	6.7	6.7	6.7	6.1	7.1	6.6	97.6	22.5	26.0	24.2	16.9	10.9	12.7	12.8	222	471	87.0	68.0	28.0	3.0	20.0	0.0	105	39.6	51.3	30.3	27.3	21.1	33.6	20.0
2204	7.3	6.7	7.3	7.3	7.3	6.8	8.0	7.1	117	28.0	23.6	23.4	15.3	10.0	14.0	14.4	106	226	36.0	34.0	15.0	4.0	13.0	2.0	155	80.3	43.2	32.1	24.8	21.2	26.2	37.4
2241	8.9	8.8	9.0	9.2	9.0	8.1	8.0	8.8	160	36.7	41.2	38.9	25.1	14.7	24.3	17.5	225	583	123	140	89.0	17.0	53.0	12.0	130	58.4	56.9	34.8	34.1	24.8	40.3	20.3
2249	6.1	5.9	6.2	6.0	6.2	5.7	6.6	6.1	92.1	15.4	23.6	19.1	11.5	11.5	9.7	9.0	136.0	212	49.0	20.0	1.0	6.0	8.0	6.0	105	30.2	41.2	26.2	20.3	24.0	23.2	25.0
Error	0.0	0.4	0.2	0.2	0.1	0.7	0.7	0.2	0.0	143	132	168	178	188	181	166	0.0	195	67.0	91.0	100	120	120	112	0.0	80.0	86.0	126	129	131.0	126	119

R and RMSE metrics combined. The null hypothesis was rejected at a 95% significance level, providing additional evidence that some models perform better than others. Having rejected the null hypothesis, the procedure stated in the previous section is followed; multiple t-tests (with the Holm correction) are carried out to verify which models present significant differences among them. The results of the multiple t-tests are shown in Table 9. It can be seen that NN performance is significantly different to the rest of the models, allowing us to conclude that NN are the best performing model for average rainfall intensity prediction.

The performance of the models reproducing some important rainfall statistics and indices, such as daily mean (μ), daily variance (σ^2), R20 (number of days with precipitation over 20 mm) and RX1 (maximum 1-day precipitation) is also evaluated (see Table 10). As with R and RMSE scores, the statistics shown in Table 10 were computed on the subset of wet days (prcp $>$ 0.1 mm). As shown in Table 10, the simulated series show BIAS values very close to 0 for all models. k-NN and SVM are the most biased ones with an average error for all stations equal to 0.7 mm/day, which corresponds to approximately 8% of rainfall intensity for the 17 gauges. The results of BIAS fall below 2% of rainfall intensity for GLM-G, NN, RF and WT methods. The σ^2, R20 and RX1 statistics are underestimated by all methods. Nevertheless, we must bear in mind that extreme events in Tenerife present values extremely high; 5 of the 17 gauges analyzed present values of RX1 above 180 mm/day during the period 1979–2015. GLM-G presents smaller errors when predicting σ^2, R20 and RX1. However, it is still below the observed values. NN performs better when reproducing the statistics σ^2, RX20 and RX1 than all other machine learning methods (with the exception of WT which presents very similar results). In contrast NN are not able to simulate events with rainfall intensities greater than 20 mm/day (see RX5 statistic in Table 10) for most of the gauges located in the northeast of the island (ID: 2010, 2112, 2115, 2118 and 2135). Most of the gauges present values of RX1 above 100 mm/day; only the statistical models (GLMs) reach that intensity for a few of the gauges. GLM-G is the method that best represent the extreme events with an average error of 86 mm/day for RX1 statistic.

Table 10 shows that simulated values of σ^2 are smaller than those observed. This is mainly because models underestimate the intensity of the most extreme values of rainfall. This effect can be appreciated more clearly when the distribution of precipitation is separated in two regimes, one for values below 20 mm/day and another for values above. Fig. 7 compares the observed and simulated distributions for rainfall intensities below 20 mm/day. Hexagons with larger number of data present darker colors. These 2D density plots allow us to appreciate that the hexagons with the largest number of data are not located in the bisector for most of the models. The figure shows that observed distributions shapes are positively skewed, with fewer data toward the larger numeric values. The mode is located close to 0.2 mm/day. Only the distribution shapes simulated by NN and SVM models mimic the observed ones. For the rest of the models, the mode is located between 2.5 mm/day and 5 mm/day.

On the other side, Fig. 8 compares the observed and simulated rainfall distribution for rainfall intensities above 20 mm/day. Red color bands mark those values not represented in the histograms (prcp $<$ 20 mm/day). It can be seen that the most extreme values are greatly underestimated by all models. Only the generalized linear models (GLM-L and GLM-G) are able to simulate values above 100 mm/day. However, as shown in Table 10, the performance for the RX1 and R20 indices is not satisfactory.

The results shown in Fig. 8 and in Table 10 demonstrate that regressors vastly underestimate the intensity of the most extreme events recorded in Tenerife during the period 1979–2015. Most of these events correspond to isolated depressions at high levels, also known as “gota fría” in Spanish. Some examples of “gota fría” include the events than happened on: 1993-03-17, 1999-01-07, 2001-03-13, 2001-11-20, 2002-03-31, 2002-12-12, 2005-12-19. If we analyze their predictors (SLP, GH500 and GH850 variables from the CFSR database) we realize that the majority of these events present very similar atmospheric synoptic patterns. However, the intensity of precipitation and their spatial distribution is completely different. This could mean: (1) that there are other explanatory variables disregarded in the analysis or (2) that the CFSR reanalysis database, with a resolution of 0.25°, does not capture the local processes that take place in Tenerife. In this context, we investigated if other explanatory variables, such as surface air temperature, relative humidity and zonal wind speed, could explain the difference in rainfall patterns for those dates; however, the atmospheric spatial patterns turned out to be also very similar. Therefore, in our opinion, the lack of spatial resolution might explain why the models are not able to predict the intensity of extreme events.

The results improve when the comparison is carried out at larger

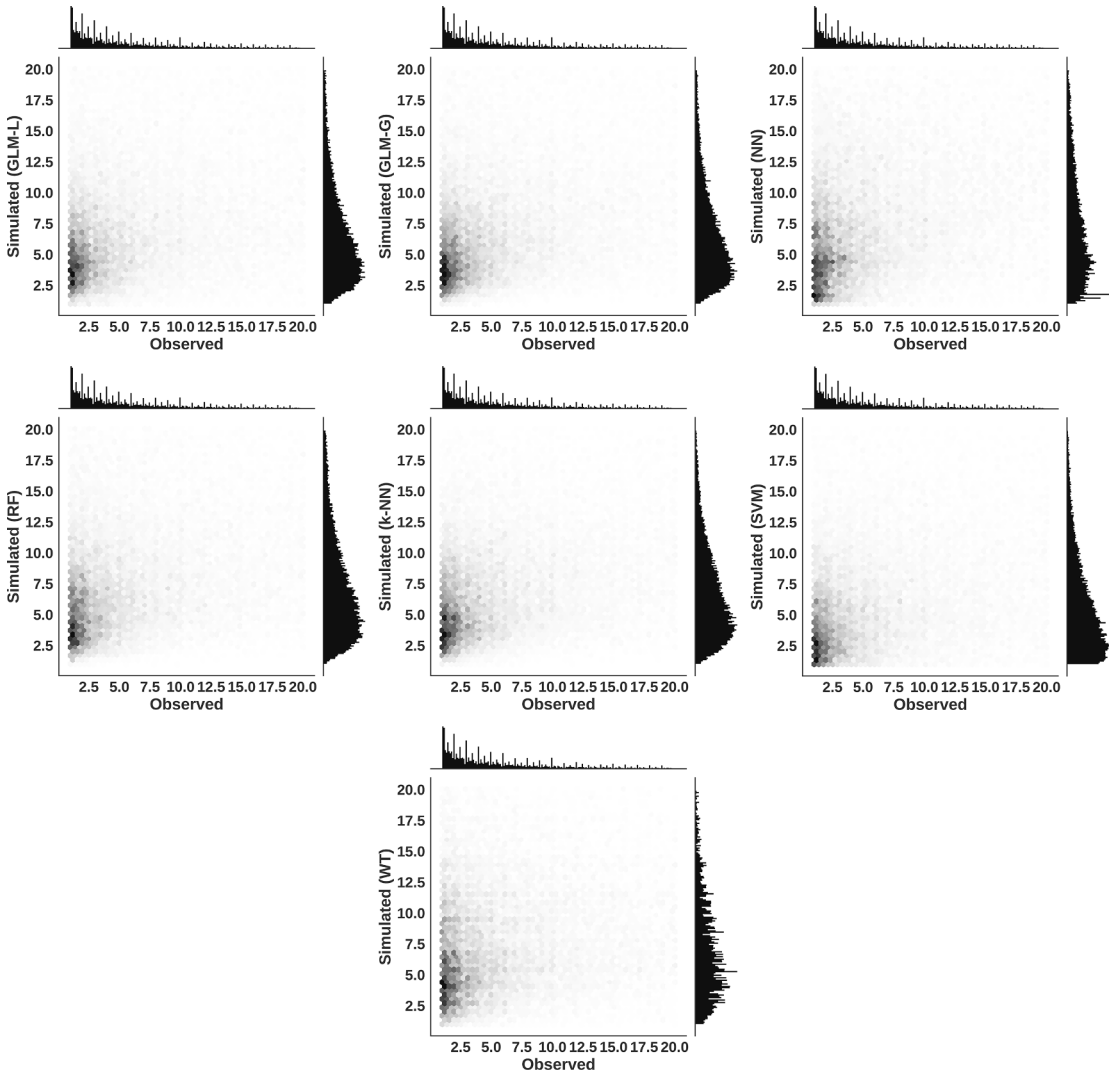

Fig. 7. Observed and simulated rainfall intensities (prcp ⩽ 20 mm/day) represented by joint probability distributions. 17 gauges are shown together. Hexagons with larger number of data present more intense colors.

temporal aggregations. Fig. 9 shows the skill of the models simulating the observed rainfall series at monthly aggregation. All models, except WT, present values of R very close to 0.8. However NN, again, presents the best skill. GLM-G and GLM-L are the only methods able to simulate months with cumulative rainfall values above 500 mm. The simulated distribution shapes reproduce adequately the observed ones, with a mode located at 50 mm and very few values above 500 mm.

Fig. 10 shows the skill of the models simulating the observed rainfall series at annual aggregation. All models show values of R very close to 0.8. k-NN and, above all, SVM, underestimate the most intense values of the distribution; they are precisely the models that show higher BIAS values (see Table 10) Ensemble methods like RF, k-NN and WT rarely simulate years with accumulated precipitation values above 800 mm.

If we investigate now the skill of the methods to predict the average rainfall intensity for the whole island (at daily scale), we appreciate that the results improve significatively with respect to those obtained for each gauge. We achieve, for example, for the NN method, a value of R equal to 0.61 when we compare observations and simulations for the 17 gauges aggregated in a single one.

Finally, it is important to note that none of the regressors used in this analysis considered location explicitly in the analysis. However, a verification was carried out to check if the spatial correlation among rainfall gauges was conserved, based only on the information provided by synoptic patterns. Fig. 11 shows the spatial correlation of the observed and the predicted rainfall at daily, monthly and annual aggregations. From left to right the panels shows the cross-correlation to a daily, monthly and annual aggregation, respectively. Panels located in the upper part of the figure show the observed spatial correlation while those located at the bottom show the quotient between simulated and observed spatial correlation for all the regressors. It can be seen that observed spatial correlation presents more pronounced curvatures at daily and annual temporal aggregations than at the monthly scale. Observed spatial correlation is slightly overestimated by most of the models (with the exception of GLM-L and WT models) for distances greater than 10 km at daily resolution. All models present successful results at the monthly scale. SVM, k-NN and RF simulate time series with higher values of spatial correlation than the observed ones at annual resolution; this happens because this methods show distributions with less variance than the observed at this resolution (see Fig. 10). Only the WT method is able to preserve the observed spatial correlation for all temporal aggregations; which is expected since it is the only method in which predictions are done simultaneously for all the gauges. The differences between the observed and simulated results, for the rest of the methods, is determined mainly because of probabilities being involved in the reconstructed time series, and pair-wise correlations not explicitly considered in the analysis.

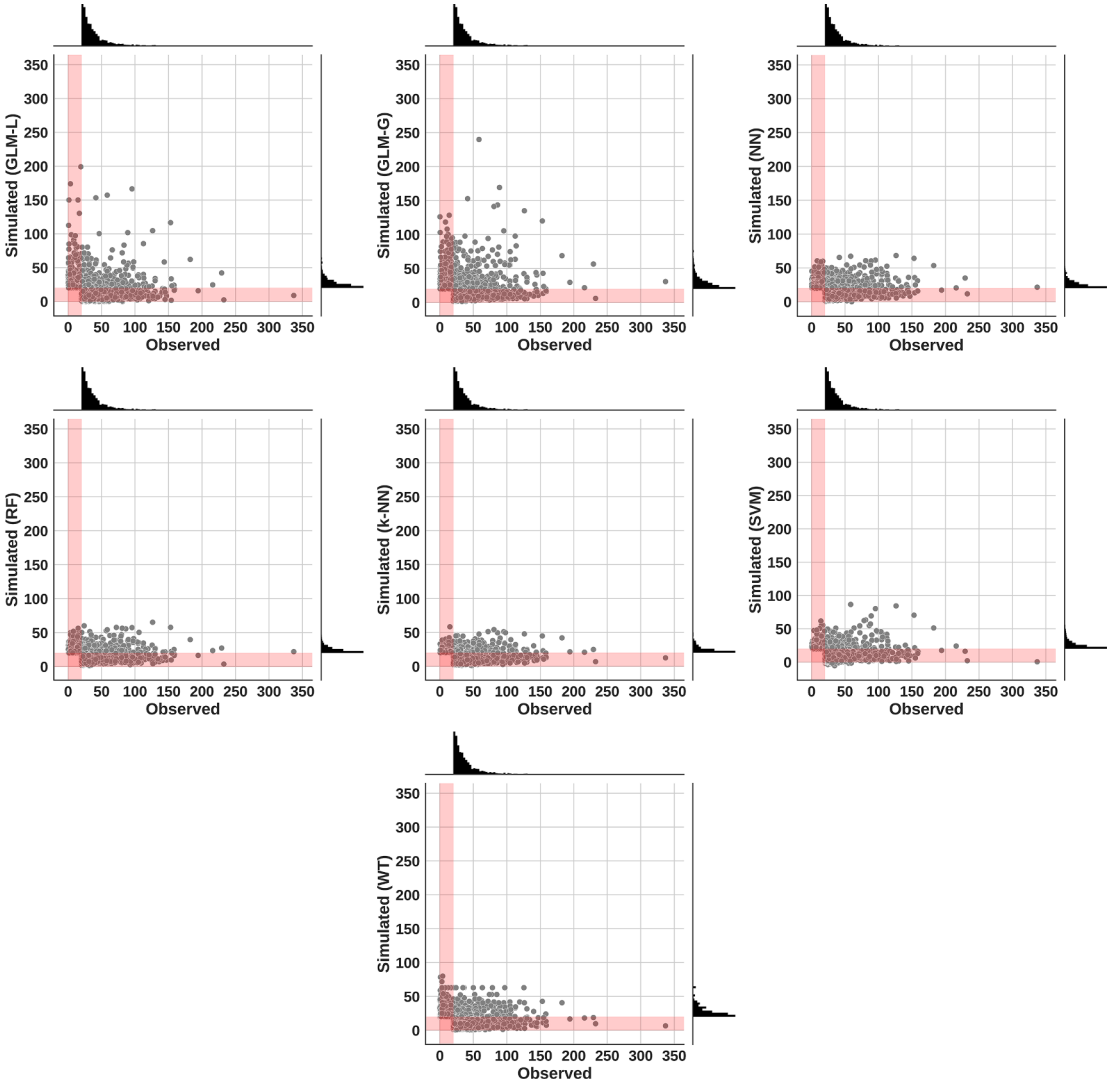

Fig. 8. Observed and simulated rainfall intensities (prcp > 20 mm/day) represented by joint probability distributions. 17 gauges are shown together. The shadowed bands mark those values not represented in the histograms.

5. Discussion and conclusions

A comparison of 8 statistical and machine learning methods for long-term rainfall prediction has been presented. All methods use atmospheric synoptic patterns for the variables SLP, GH500 and GH850 as predictors. The analysis was carried out for 17 gauges along the island of Tenerife (Spain), that presents a semi-arid climate. The quality of the predictions was evaluated using different metrics on rainfall occurrence and on rainfall intensity comparing modeled and observed values at different temporal aggregations by cross-validation.

We show that most machine learning methods are very sensitive to the hyperparameters chosen. Wrong parameterization could lead to models without any predictive capability. It may also lead to models that overfit the training data. RF and k-NN methods tend to over-fit. In order to avoid over-fitting, the hyperparameters *N_estimators* and *min_samples_leaf*, in the case of RF, and *n_neighbors* for k-NN should take higher values, even when doing so degrades their training skill. Hyperparameters that result in the best predictions changed considerably from one rainfall station to the others. Our work demonstrates that it is essential to determine the optimal hyperparameters before any application, instead of resourcing to any predefined recommendation. To minimize the computational cost, we recommend starting with a wide search range, and then, center the search around the hyperparameters that offered better results in the first search.

The results show that NN, with an average f-score close to 0.4 for daily rainfall prediction, is the best method for predicting rainfall occurrence, closely followed by LR and SVM. All models are able to preserve ϕ_{DD}, however they underestimate the observed values of ϕ_{WW}. NN also presents less error when predicting the transition probabilities, mainly due to its high skill predicting rainfall occurrence. The increased in recall metric when prediction threshold is raised demonstrates that more intense rainfall events are easier to predict.

All models except WT reach values of R above 0.8 when tested simulating the number of wet days per month and per year. NN, SVM and LR, however, are the only methods able to simulate months with more than 20 wet days, which is exceptionally unusual in Tenerife.

NN presents significantly better results than the rest of the models when predicting the intensity of rainfall. SVM, k-NN and RF rank second, with slightly worse values of R and RMSE than NN. WT presents the worst results. Only rainfall intensity distribution shapes simulated by NN and SVM models mimic the observed ones. Simulated series show BIAS values very close to 0 for all models.

All methods underestimate the variance of the observed series. They are not able to simulate events with accumulated daily rainfall values as high as the observed ones. GLM-G is the model that best reproduces the extreme indices RX1 and R20, however, the results are still

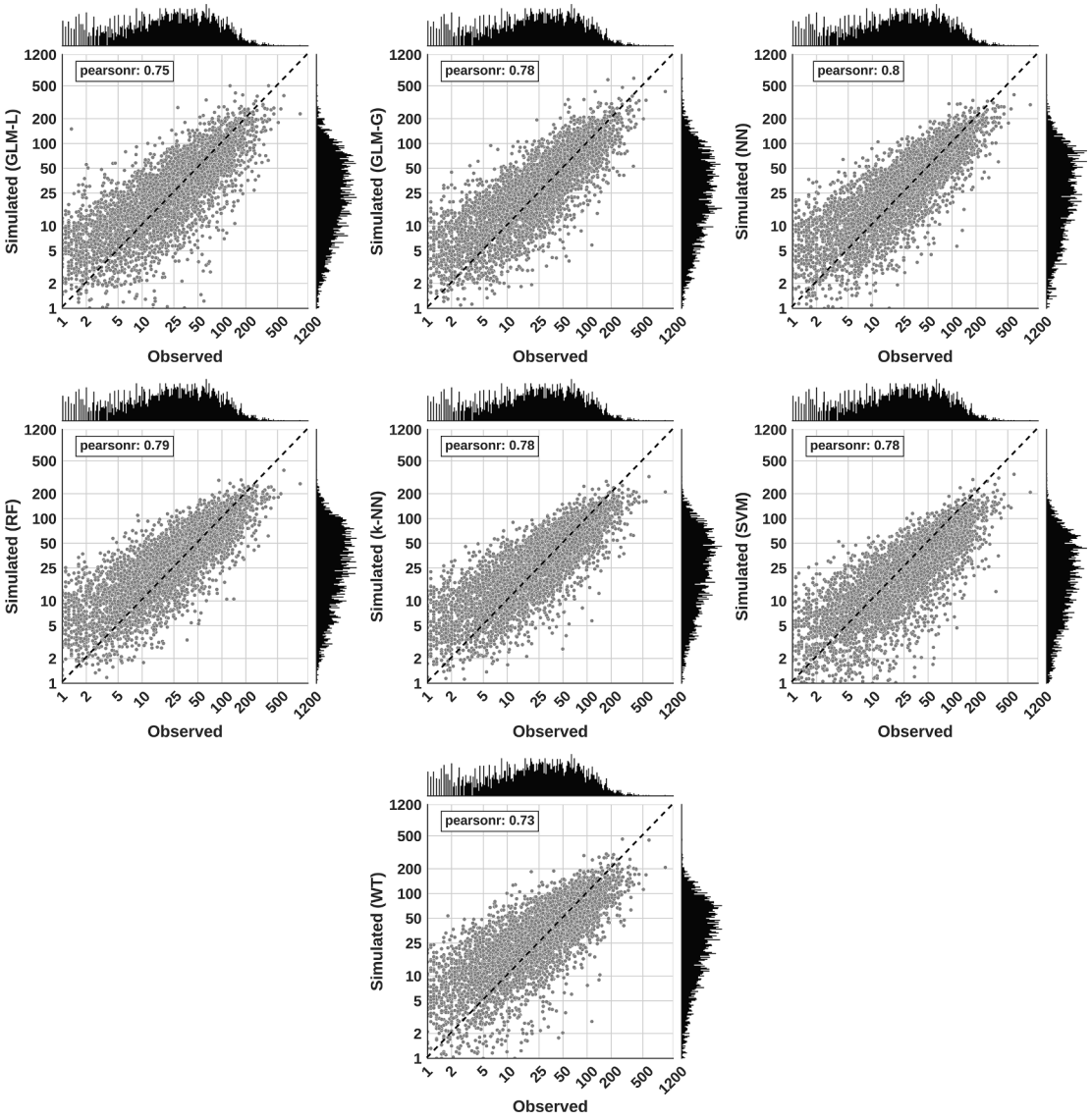

Fig. 9. Observed and simulated monthly rainfall amounts represented by joint probability distributions. 17 gauges are shown together. Dashed line indicates a perfect fit.

improvables. Most of the extreme events correspond to isolate high altitude depression, also known as “gota fría” in Spanish. We realized that, in these situations, the atmospheric synoptic patterns for variables SLP, GH500 and GH850 were indistinguishable, however, the intensity of the precipitation and their spatial distribution were completely different. It might indicate that local processes, in Tenerife Island, are not well capture by the CFSR database.

When comparing the observed and simulated series for larger temporal aggregations, the skill of the methods improve significantly. All models, except WT, present values of R close to 0.8 predicting rainfall intensity. NN, again, presents the best skill. GLM-G is the only method able to simulate months with cumulative rainfall values above 500 mm. Ensemble methods like RF, k-NN and WT rarely simulate years with accumulated precipitation values above 800 mm.

Though regression is done independently for each gauge, the atmospheric predictors allow most of the models to preserve the observed spatial correlation at daily, monthly and annual aggregations. A small error appears for distances greater than 10 km at daily scale. The methods that show less variance at an annual temporal aggregation (SVM, k-NN and RF) simulate series with higher values of spatial correlation than the observed ones. Only the WT method is able to preserve the observed spatial correlation for all temporal aggregations; which is expected since it is the only method in which predictions are done for all the gauges simultaneously. The incorporation of conceptual spatio-temporal rainfall models using copulas, or deep learning techniques (Stehlík and Bárdossy, 2002; Yang et al., 2005) will be considered in future works.

Our results are in line with other studies performed in the past (Gupta and Ghose, 2015; Olsson et al., 2004; Valverde Ramírez et al., 2005), however, none of these studies carried out a significance analysis to demonstrate the best predictive capacity of some methods against others. Gupta and Ghose (2015) found that NN provides better results than any of the other discussed algorithms (Decision tree, Naive Bayes approach, K-Nearest Neighbor) for predicting rainfall occurrence in New Delhi from June to September (rainfall period). Weather predictors including mean temperature, dew point temperature, humidity, sea level pressure and wind speed. According to the results achieved by Gupta and Ghose (2015), NN achieved a value of f-score equal to 0.6.

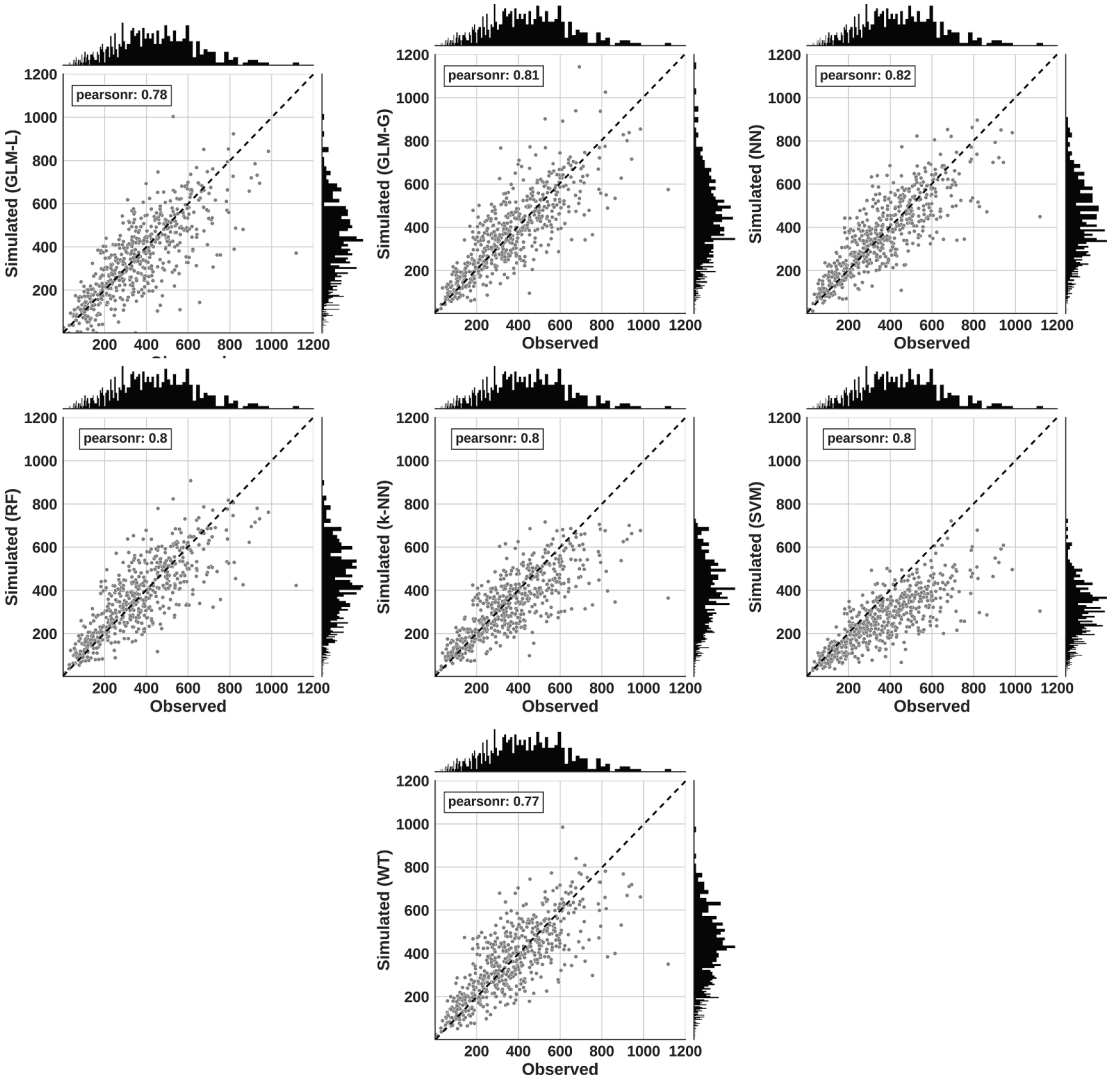

Fig. 10. Observed and simulated annual rainfall amounts represented by joint probability distributions. 17 gauges are shown together. Dashed line indicates a perfect fit.

However their analysis was carried out only at a single location (gauge) and, furthermore, methods were fitted and validated only during rainfall period, which probably facilitates the prediction. In our analysis we obtain an average f-score of 0.4 for the 17 gauges, however, it can vary from 0.24 to 0.54 depending on the gauge. Olsson et al. (2004) used NN for predicting 12-h mean rainfall in Kyushu Island, southern Japan, from values of wind speeds at 850 hPa and precipitable water in a 100X100 km grid surrounding the island. In their study, Olsson et al. (2004) employ a classification of intensities into four categories (zero, low, high and extreme intensity). They achieved a result of R equal to 0.63 for the entire island. In our analysis we achieve an average result of 0.37 for the 17 gauges. We have to bear in mind that in their analysis, Olsson et al. (2004) use atmospheric instrumental data from the gauges. We instead use large-scale atmospheric reanalysis data. It is expected that including atmospheric instrumental data in our analysis will improve the predictions. In addition, Olsson et al. (2004) predict the average precipitation in a cell of 100X100 km. If we compare their results with those obtained when we aggregate rainfall intensity for all the gauges in the study area, we realize that the results are very similar. In our case we obtained a value of R closet to 0.61. Valverde Ramírez et al. (2005) compared NN with a multiple linear regression model to predict daily rainfall using as predictors output data from a regional climatic model. They use the variables potential temperature, vertical component of the wind, specific humidity, air temperature, precipitable water, relative vorticity and moisture divergence flux in their analysis. The test was performed for six locations in São Paulo State, Brazil, during the austral summer and winter for the period 1997–2002. The results show that NN forecasts were superior to the ones obtained by the linear regression model. NN achieves results of R between 0.1 and 0.8, depending on the month and gauge analyzed. The results obtained in our article also indicate that NN is the best method to predict rainfall occurrence and rainfall intensity at the daily scale. However, to verify this hypothesis, we carried out a multivariate and univariate analysis of variance to evaluate the difference among models. Furthermore, we demonstrate that none of the models is able to reproduce the variance and the intensity of the highest values of the distribution, which suggests that these methods are not appropriate to reproduce the extreme events that took place in the past. The underestimation of variance could have been addressed by including aggregate predictors in the simulation context or by using nonparametric methods based on kernel density estimation (Mehrotra et al., 2006; Mehrotra et al., 2004; Sharma and Lall, 1999), by using mixed distributions (Mínguez et al., 2013; Jeffries and Pfeiffer, 2001) or by testing other minimization function; all these approaches will be investigated in future works. The good results achieved at temporal aggregations above the day indicate that these methods may be very useful tools for water resources studies.

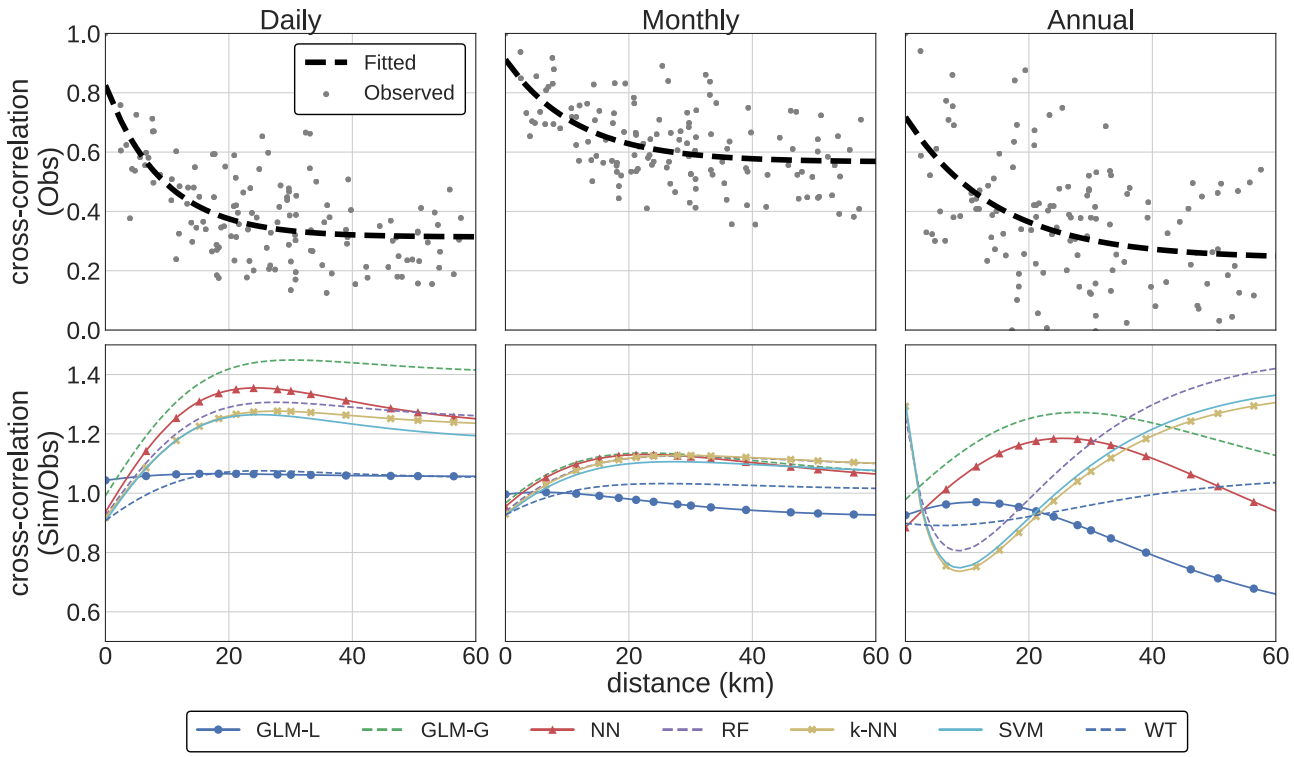

Fig. 11. Observed and simulated spatial cross-correlation for all the regressors. From left to right the panels show the cross-correlation to a daily, monthly and annual aggregation, respectively. Panels located in the upper part of the figure show the observed spatial correlation while those located at the bottom show the quotient between simulated and observed spatial correlation for all the regressors.

CRediT authorship contribution statement

Javier Diez-Sierra: Methodology, Investigation, Visualization. **Manuel del Jesus:** Conceptualization, Methodology, Supervision, Funding acquisition.

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 would like to thank: (1) “Consejo Insular de Aguas de Tenerife (CIATF)” for granting permission to use their rainfall database for this work; (2) “Agencia Estatal de Investigación (AEI)” from the Spanish Ministry of Economy, Industry and Competitiveness, and the European Regional Development Fund (ERDF) for the funding provided through grant BIA2016-78397-P (AEI/FEDER, UE); and (3) Project INDECIS, which is part of ERA4CS, an ERA – NET initiated by JPI Climate, and funded by FORMAS (SE), DLR (DE), BMWFW (AT), IFD (DK), MINECO (ES), ANR (FR) with co – funding by the European Union (Grant 690462).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jhydrol.2020.124789.

References

Abbot, J., Marohasy, J., 2017. Forecasting of Medium-term Rainfall Using Artificial Neural Networks: Case Studies from Eastern Australia, Engineering and Mathematical Topics in Rainfall, Theodore V Hromadka II and Prasada Rao. IntechOpen.

Abdi, H., Williams, L.J., 2010. Principal component analysis. Wiley Interdiscip. Rev. 2 (4), 433–459. https://doi.org/10.1002/wics.10. ISSN 1939-0068.

Adeyewa, Z., Nakamura, K., 2003. Validation of TRMM radar rainfall data over major climatic regions in Africa. J. Appl. Meteorol. 42 (2), 331–347. https://doi.org/10.1175/1520-0450(2003)042<0331:VOTRRD>2.0.CO;2. cited By 124.

Adnan, R.M., Liang, Z., Trajkovic, S., Zounemat-Kermani, M., Li, B., Kisi, O., 2019. Daily streamflow prediction using optimally pruned extreme learning machine. J. Hydrol. 577, 123981. https://doi.org/10.1016/j.jhydrol.2019.123981. ISSN 0022-1694.

AEMET, 2018. Agencia Estatal de Meteorología. URL:http://www.aemet.es/en/portada (accessed: 2018-07-04).

AgroCabildo, 2018. Servicio Técnico de Agricultura y Desarrollo Rural. URL:http://www.agrocabildo.org/ (accessed: 2018-07-04).

Altunkaynak, A., Nigussie, T., 2015. Prediction of daily rainfall by a hybrid wavelet-season-neuro technique. J. Hydrol. 529 (P1), 287–301. https://doi.org/10.1016/j.jhydrol.2015.07.046.

Appelhans, T., Mwangomo, E., Hardy, D., Hemp, A., Nauss, T., 2015. Evaluating machine learning approaches for the interpolation of monthly air temperature at Mt. Kilimanjaro, Tanzania. Spatial Stat. 14, 91–113. https://doi.org/10.1016/j.spasta.2015.05.008.

Aryal, Y.N., Villarini, G., Zhang, W., Vecchi, G.A., 2018. Long term changes in flooding and heavy rainfall associated with North Atlantic tropical cyclones: roles of the North Atlantic Oscillation and El Niño-Southern Oscillation. J. Hydrol. 559, 698–710. https://doi.org/10.1016/j.jhydrol.2018.02.072. ISSN 0022-1694.

Austin, G., Seed, A., 2005. Special issue on the hydrological applications of weather radar – guest editors’ preface. Atmos. Sci. Lett. 6 (1), 1.

Baarsch, J., Celebi, M., 2012. Investigation of internal validity measures for K-means clustering. 2195, 471–476.

Ben Alaya, M., Ouarda, T., Chebana, F., 2017. Non-Gaussian spatiotemporal simulation of multisite daily precipitation: downscaling framework. Clim. Dyn. 1–15. https://doi.org/10.1007/s00382-017-3578-0.

Bosilovich, M., Chen, J., Robertson, F., Adler, R., 2008. Evaluation of global precipitation in reanalyses. J. Appl. Meteorol. Climatol. 47 (9), 2279–2299. https://doi.org/10.1175/2008JAMC1921.1. cited By 208.

Burlando, P., Montanari, A., Ranzi, R., 1996. Forecasting of storm rainfall by combined use of radar, rain gages and linear models. Atmos. Res. 42 (1–4), 199–216. https://doi.org/10.1016/0169-8095(95)00063-1. cited By 15.

Buytaert, W., Friesen, J., Liebe, J., Ludwig, R., 2012. Assessment and management of water resources in developing, semi-arid and arid regions. Water Resour. Manage. 26 (4), 841–844. https://doi.org/10.1007/s11269-012-9994-3. cited By 13.

Camus, P., Mendez, F., Medina, R., Cofiño, A., 2011. Analysis of clustering and selection algorithms for the study of multivariate wave climate. Coast. Eng. 58 (6), 453–462. https://doi.org/10.1016/j.coastaleng.2011.02.003.

Chen, H., Yong, B., Shen, Y., Liu, J., Hong, Y., Zhang, J., 2020. Comparison analysis of six purely satellite-derived global precipitation estimates. J. Hydrol. 581, 124376.

https://doi.org/10.1016/j.jhydrol.2019.124376. ISSN 0022-1694.
Cherkassky, V., Ma, Y., 2004. Practical selection of SVM parameters and noise estimation for SVM regression. Neural Networks 17 (1), 113–126.
CIATF, Consejo Insular de Aguas de Tenerife, 2018. URL:http://www.aguastenerife.org/ (accessed 2018-07-04).
Coe, R., Stern, R., 1982. Fitting models to daily rainfall data. 21, 1024–1031.
Dee, D., Uppala, S., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hersbach, H., Hólm, E., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., Mcnally, A., Monge-Sanz, B., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., Vitart, F., 2011. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Q. J. R. Meteorol. Soc. 137 (656), 553–597. https://doi.org/10.1002/qj.828. cited By 12344.
Del Jesus, M., Rinaldo, A., Rodríguez-Iturbe, I., 2015. Point rainfall statistics for ecohydrological analyses derived from satellite integrated rainfall measurements. Water Resour. Res. 51 (4), 2974–2985.
Diez-Sierra, J., del Jesus, M., 2017. A rainfall analysis and forecasting tool. Environ. Modell. Software 97, 243–258. https://doi.org/10.1016/j.envsoft.2017.08.011.
Diez-Sierra, J., del Jesus, M., 2019. Subdaily rainfall estimation through daily rainfall downscaling using random forests in Spain. Water (Switzerland) 11 (1). https://doi.org/10.3390/w11010125.
D. Gupta, U. Ghose, A comparative study of classification algorithms for forecasting rainfall, 2015.https://doi.org/10.1109/ICRITO.2015.7359273.
Gutiérrez, J., Cofino, A., Cano, R., Rodríguez, M., 2004. Clustering methods for statistical downscaling in short-range weather forecasts. Mon. Weather Rev. 132 (9), 2169–2183. https://doi.org/10.1175/1520-0493(2004)132<2169:CMFSDI>2.0.CO;2.
Gutiérrez, J.M., San-Martín, D., Brands, S., Manzanas, R., Herrera, S., 2013. Reassessing statistical downscaling techniques for their robust application under climate change conditions. J. Clim. 26 (1), 171–188. https://doi.org/10.1175/JCLI-D-11-00687.1.
Hasan, M., Sharma, A., Johnson, F., Mariethoz, G., Seed, A., 2016. Merging radar and in situ rainfall measurements: an assessment of different combination algorithms. Water Resour. Res. 52 (10), 8384–8398. https://doi.org/10.1002/2015WR018441. cited By 13.
He, X., Guan, H., Qin, J., 2015. A hybrid wavelet neural network model with mutual information and particle swarm optimization for forecasting monthly rainfall. J. Hydrol. 527, 0022–1694.
Herrera, R.G., Puyol, D.G., Martín, E.H., Presa, L.G., Rodríguez, P.R., 2001. Influence of the North Atlantic oscillation on the Canary Islands precipitation. J. Clim. 14 (19), 3889–3903.
Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scand. J. Stat. 6, 65–70.
Hong, W.-C., 2008. Rainfall forecasting by technological machine learning models. Appl. Math. Comput. 200 (1), 41–57. https://doi.org/10.1016/j.amc.2007.10.046.
G. Huffman, R. Adler, D. Bolvin, E. Nelkin, The TRMM Multi-satellite Precipitation Analysis (TMPA), 2010.
Jeffries, N., Pfeiffer, R., 2001. A mixture model for the probability distribution of rain rate. Environmetrics 12 (1), 1–10.
Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K., Ropelewski, C., Wang, J., Leetmaa, A., Reynolds, R., Jenne, R., Joseph, D., 1996. The NCEP/NCAR 40-year reanalysis project. Bull. Am. Meteorol. Soc. 77 (3), 437–471. https://doi.org/10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2. cited By 20383.
Kannan, S., Ghosh, S., 2013. A nonparametric kernel regression model for downscaling multisite daily precipitation in the Mahanadi basin. Water Resour. Res. 49 (3), 1360–1385. https://doi.org/10.1002/wrcr.20118.
Li, Y., Wang, Q., He, H., Wu, Z., Lu, G., 2020. A method to extend temporal coverage of high quality precipitation datasets by calibrating reanalysis estimates. J. Hydrol. 581, 124355. https://doi.org/10.1016/j.jhydrol.2019.124355. ISSN 0022-1694.
Markatou, M., Tian, H., Biswas, S., Hripcsak, G., 2005. Analysis of variance of cross-validation estimators of the generalization error. J. Mach. Learn. Res. 6.
Mattera, D., Haykin, S., 1999. Advances in Kernel Methods, Chap. Support Vector Machines for Dynamic Reconstruction of a Chaotic System. MIT Press, Cambridge, MA.
Mehrotra, R., Sharma, A., Cordery, I., 2004. Comparison of two approaches for downscaling synoptic atmospheric patterns to multisite precipitation occurrence. J. Geophys. Res.: Atmospheres 109 (D14). https://doi.org/10.1029/2004JD00482.
Mehrotra, R., Srikanthan, R., Sharma, A., 2006. A comparison of three stochastic multi-site precipitation occurrence generators. J. Hydrol. 331 (1), 280–292. https://doi.org/10.1016/j.jhydrol.2006.05.016. ISSN 0022-1694. Water Resources in Regional Development: The Okavango River.
Melián, P.D., Ruiz, J.-J.B., Díez, A.M., Salete, E.G., 2011. Módulo de Gestión de Tormentas" en la modelización hidrológica de superficie de Tenerife. In: II Jornadas de Ingeniería del Agua, Fundación para el Fomento de la Ingeniería del Agua, Barcelona, Spain.
Méndez, F., Menéndez, M., Luceño, A., Losada, I., 2007. Analyzing monthly extreme sea levels with a time-dependent GEV model. J. Atmos. Ocean. Technol. 24 (5), 894–911. https://doi.org/10.1175/JTECH2009.1.
Mínguez, R., Tomás, A., Méndez, F., Medina, R., 2013. Mixed extreme wave climate model for reanalysis databases. Stoch. Environ. Res. Risk Assess. 27 (4), 757–768. https://doi.org/10.1007/s00477-012-0604-y.
Muller, K., Peterson, B., 1984. Practical methods for computing power in testing the multivariate general linear hypothesis. Comput. Stat. Data Anal. 2 (2), 143–158. https://doi.org/10.1016/0167-9473(84)90002-1.
Nkiaka, E., Nawaz, N.R., Lovett, J.C., 2017. Evaluating global reanalysis precipitation datasets with rain gauge measurements in the Sudano-Sahel region: case study of the Logone catchment, Lake Chad Basin. Meteorol. Appl. 24 (1), 9–18. https://doi.org/10.1002/met.1600.
Nkiaka, E., Nawaz, N., Lovett, J., 2017. Evaluating global reanalysis precipitation datasets with rain gauge measurements in the Sudano-Sahel region: case study of the Logone catchment, Lake Chad Basin. Meteorol. Appl. 24 (1), 9–18. https://doi.org/10.1002/met.1600. cited By 12.
Olsson, J., Uvo, C., Jinno, K., Kawamura, A., Nishiyama, K., Koreeda, N., Nakashima, T., Morita, O., 2004. Neural networks for rainfall forecasting by atmospheric downscaling. J. Hydrol. Eng. 9 (1), 1–12. https://doi.org/10.1061/(ASCE)1084-0699(2004)9:1(1).
Park, S., Berenguer, M., Sempere-Torres, D., 2019. Long-term analysis of gauge-adjusted radar rainfall accumulations at European scale. J. Hydrol. 573, 768–777. https://doi.org/10.1016/j.jhydrol.2019.03.093. ISSN 0022-1694.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Pérez-Rodríguez, P., Gianola, D., González-Camacho, J., Crossa, J., Manès, Y., Dreisigacker, S., 2012. Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. G3: Genes Genomes Genet. 2 (12), 1595–1605. https://doi.org/10.1534/g3.112.003665.
Pfeifroth, U., Mueller, R., Ahrens, B., 2013. Evaluation of satellite-based and reanalysis precipitation data in the tropical pacific. J. Appl. Meteorol. Climatol. 52 (3), 634–644.
Preisendorfer, R., 1988. Principal component analysis in meteorology and oceanography, cited By 1183.
Pumo, D., Arnone, E., Francipane, A., Caracciolo, D., Noto, L., 2017. Potential implications of climate change and urbanization on watershed hydrology. J. Hydrol. 554, 80–99. https://doi.org/10.1016/j.jhydrol.2017.09.002.
Qiu, J., Wu, Q., Ding, G., Xu, Y., Feng, S., 2016. A survey of machine learning for big data processing. Eurasip J. Adv. Signal Process. 1. https://doi.org/10.1186/s13634-016-0355-x.
R Core Team, 2017. R A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL:https://www.R-project.org/.
Saha, S., Moorthi, S., Pan, H.-L., Wu, X., Wang, J., Nadiga, S., Tripp, P., Kistler, R., Woollen, J., Behringer, D., Liu, H., Stokes, D., Grumbine, R., Gayno, G., Wang, J., Hou, Y.-T., Chuang, H.-Y., Juang, H.-M., Sela, J., Iredell, M., Treadon, R., Kleist, D., Van Delst, P., Keyser, D., Derber, J., Ek, M., Meng, J., Wei, H., Yang, R., Lord, S., Van Den Dool, H., Kumar, A., Wang, W., Long, C., Chelliah, M., Xue, Y., Huang, B., Schemm, J.-K., Ebisuzaki, W., Lin, R., Xie, P., Chen, M., Zhou, S., Higgins, W., Zou, C.-Z., Liu, Q., Chen, Y., Han, Y., Cucurull, L., Reynolds, R., Rutledge, G., Goldberg, M., 2010a. The NCEP climate forecast system reanalysis. Bull. Am. Meteorol. Soc. 91 (8), 1015–1057. https://doi.org/10.1175/2010BAMS3001.1, cited By 2622.
Saha, S., Moorthi, S., Pan, H.-L., Wu, X., Wang, J., Nadiga, S., Tripp, P., Kistler, R., Woollen, J., Behringer, D., et al., 2010b. The NCEP climate forecast system reanalysis. Bull. Am. Meteorol. Soc. 91 (8), 1015–1057.
San-Martín, D., Manzanas, R., Brands, S., Herrera, S., Gutiérrez, J.M., 2017. Reassessing model uncertainty for regional projections of precipitation with an ensemble of statistical downscaling methods. J. Clim. 30 (1), 203–223. https://doi.org/10.1175/JCLI-D-16-0366.1. cited By 16.
Scikit-learn, 2019. Scikit-learn. URL:https://scikit-learn.org/stable/ [Online; accessed 22-August-2019].
Serrano-Notivoli, R., Beguería, S., Saz, M., de Luis, M., 2018. Recent trends reveal decreasing intensity of daily precipitation in Spain. Int. J. Climatol. 38 (11), 4211–4224. https://doi.org/10.1002/joc.5562.
Sharma, A., Lall, U., 1999. A nonparametric approach for daily rainfall simulation. Math. Comput. Simul. 48 (4), 361–371. https://doi.org/10.1016/S0378-4754(99)00016-6. ISSN 0378-4754.
Sharma, A., Mehrotra, R., Li, J., Jha, S., 2016. A programming tool for nonparametric system prediction using Partial Informational Correlation and Partial Weights. Environ. Modell. Software 83, 271–275. https://doi.org/10.1016/j.envsoft.2016.05.021, cited By 15.
Sheskin, D., 2003. Handbook of Parametric and Nonparametric Statistical Procedures, third ed. CRC Press ISBN 9781420036268.
Stehlík, J., Bárdossy, A., 2002. Multivariate stochastic downscaling model for generating daily precipitation series based on atmospheric circulation. J. Hydrol. 256 (1–2), 120–141. https://doi.org/10.1016/S0022-1694(01)00529-7.
Stephenson, D., Rupa Kumar, K., Doblas-Reyes, F., Royer, J.-F., Chauvin, F., Pezzulli, S., 1999. Extreme daily rainfall events and their impact on ensemble forecasts of the Indian monsoon. Mon. Weather Rev. 127 (9), 1954–1966.
Stern, R.D., Coe, R., 1984. A model fitting analysis of daily rainfall data. J. R. Stat. Soc. Ser. A (General) 147 (1) pp. 1–34, ISSN 00359238.
Sumi, S., Zaman, M., Hirose, H., 2012. A rainfall forecasting method using machine learning models and its application to the fukuoka city case. Int. J. Appl. Math. Comput. Sci. 22 (4), 841–854. https://doi.org/10.2478/v10006-012-0062-1.
Sun, Q., Miao, C., Duan, Q., Ashouri, H., Sorooshian, S., Hsu, K.-L., 2018. A review of global precipitation data sets: data sources, estimation, and intercomparisons. Rev. Geophys. 56 (1), 79–107. https://doi.org/10.1002/2017RG000574. cited By 144.
Tullot, I.F., 1959. El clima de las Islas Canarias. Anuario de estudios atlánticos 1 (5), 57–103.
Valverde Ramírez, M., De Campos Velho, H., Ferreira, N., 2005. Artificial neural network technique for rainfall forecasting applied to the São Paulo region. J. Hydrol. 301 (1–4), 146–162. https://doi.org/10.1016/j.jhydrol.2004.06.028.

Wang, G., Zhang, X., Zhang, S., 2019. Performance of three reanalysis precipitation datasets over the qinling-daba mountains, eastern fringe of tibetan plateau, China. Adv. Meteorol. https://doi.org/10.1155/2019/7698171. cited By 0.

Yang, C., Chandler, R., Isham, V., Wheater, H., 2005. Spatial-temporal rainfall simulation using generalized linear models. Water Resour. Res. 41 (11), 1–13. https://doi.org/10.1029/2004WR003739.

Yu, P.-S., Yang, T.-C., Chen, S.-Y., Kuo, C.-M., Tseng, H.-W., 2017. Comparison of random forests and support vector machine for real-time radar-derived rainfall forecasting. J. Hydrol. 552, 92–104 ISSN 0022-1694.

Zhao, G., Pang, B., Xu, Z., Xu, L., 2020. A hybrid machine learning framework for real-time water level prediction in high sediment load reaches. J. Hydrol. 581, 124422. https://doi.org/10.1016/j.jhydrol.2019.124422. ISSN 0022-1694.