atmosphere

MDPI

Article

Assessing Machine Learning Models for Gap Filling Daily Rainfall Series in a Semiarid Region of Spain

Juan Antonio Bellido-Jiménez *, Javier Estévez Gualda and Amanda Penélope García-Marín

Projects Engineering Area, Department of Rural Engineering, University of Córdoba, 14071 Córdoba, Spain; jestevez@uco.es (J.E.G.); amanda.garcia@uco.es (A.P.G.-M.)
* Correspondence: p22bejij@uco.es

Abstract: The presence of missing data in hydrometeorological datasets is a common problem, usually due to sensor malfunction, deficiencies in records storage and transmission, or other recovery procedures issues. These missing values are the primary source of problems when analyzing and modeling their spatial and temporal variability. Thus, accurate gap-filling techniques for rainfall time series are necessary to have complete datasets, which is crucial in studying climate change evolution. In this work, several machine learning models have been assessed to gap-fill rainfall data, using different approaches and locations in the semiarid region of Andalusia (Southern Spain). Based on the obtained results, the use of neighbor data, located within a 50 km radius, highly outperformed the rest of the assessed approaches, with RMSE (root mean squared error) values up to 1.246 mm/day, MBE (mean bias error) values up to −0.001 mm/day, and R^2 values up to 0.898. Besides, inland area results outperformed coastal area in most locations, arising the efficiency effects based on the distance to the sea (up to an improvement of 63.89% in terms of RMSE). Finally, machine learning (ML) models (especially MLP (multilayer perceptron)) notably outperformed simple linear regression estimations in the coastal sites, whereas in inland locations, the improvements were not such significant.

Keywords: gap-filling; rainfall series; machine learning; Bayesian optimization

Citation: Bellido-Jiménez, J.A.; Gualda, J.E.; García-Marín, A.P. Assessing Machine Learning Models for Gap Filling Daily Rainfall Series in a Semiarid Region of Spain. *Atmosphere* **2021**, *12*, 1158. https://doi.org/10.3390/atmos12091158

Academic Editor: Tomeu Rigo

Received: 21 August 2021
Accepted: 6 September 2021
Published: 9 September 2021

1. Introduction

The spatial and temporal analysis of meteorological parameters, such as rainfall is crucial to numerous environmental, hydrological, and agroclimatic studies, as well as optimizing issues, such as water resource management or irrigation scheduling [1–4]. However, one of the most common problems in time series analyses, such as rainfall datasets, is the presence of gaps of different widths, making this task harder to carry out. This usually results from malfunctioning sensors or data loggers, lack of maintenance, meteorological events, or power outages. Sometimes, the solution is not instantaneous and causes delays because it needs the interaction of qualified personnel. Therefore, before starting with analyses, a common practice is to fill these gaps using different methodologies and applying automatic detection algorithms to detect spurious signals in automated weather stations [5].

Due to its high spatio-temporal variability and the large number of interconnected variables involved, rainfall is one of the most challenging atmospheric variables to characterize, estimate, and forecast [6], especially on a daily basis, with higher volatility and chaotic patterns [7]. A variety of techniques have been developed on both a monthly and daily basis. One of the most frequent algorithms to estimate missing rainfall records is the inverse distance weighting method (IDWM), where the estimated values are calculated with a weighted average (it resorts to the inverse of the distance when assigning the weights) from neighbor stations [8,9]. Another simple method to apply is the gauge mean estimator, which uses an average value of observations from the nearby stations, which can be obtained by optimization, proximity metric, or correlation, among other techniques [10].

Ordinary kriging is a spatially-dependent variance, based on scalar measurements at different locations, where the weights are derived from the distance between the source and the target stations [11–13]. However, these three methods tend to overestimate the number of rainy days and underestimate their magnitudes, and even a negative correlation is found in several reports between close stations [13,14]. Xia et al. [15] evaluated six methodologies (simple arithmetic averaging, inverse distance interpolation, normal ratio method, single best estimator, multiple regression analysis (REG), and closest station method) for estimating missing data in two stations in Germany and concluded that REG consistently obtained the best performance. Teegavarapu and Chandramouli [8] highlighted that the use of the coefficient of correlation provided an improvement in estimating the missing data and recommended the coefficient of correlation weighing method, artificial neural network estimation method, and kriging estimation method for this purpose, due to their conceptually superior performance. Teegavarapu et al. [16] introduced the fixed functional set genetic algorithm method (FFSGAM), eliminating the use of rigid functional forms and weighting approaches for gap-filling. FFSGAM outperformed conventional IDWM. Adhikary et al. [12] developed genetic programming-based ordinary kriging (GPOK) as a new variant of the kriging method, using the genetic programing-derived variogram model and ordinary kriging. GPOK obtained the best results, when compared to ANN-based ordinary kriging and traditional ordinary kriging. Different authors [17–19] have evaluated the k-nearest-neighbor algorithm, in conjunction with machine learning models, such as multilayer perceptron (MLP), support vector machine (SVM), and random forest (RF), with promising results. Bagirov et al. [20] evaluated clusterwise linear regression (CLR), using different combinations of maximum and minimum daily air temperature, evaporation, vapor pressure, and solar radiation to predict monthly rainfall in Victoria, Australia. Their results showed a higher performance of CLR against different methods, such as cluster regression-expectation maximization, multiple linear regression, support vector regression (SVR), and MLP. Kajewska-Szkudlarek [21] assessed the use of cluster analysis with SVR to outperform daily rainfall prediction in urban areas.

Additionally, other researchers study the performance of processing algorithms, such as wavelets [22,23], variational mode decomposition (VMD) [24], or singular spectrum analysis (SSA) [25,26]. Estévez et al. [22] evaluated different combinations of wavelet analysis with thermo-pluviometric variables, using MLP in sixteen locations of Spain to forecast monthly rainfall. The results indicated the suitability of the models using thermo-pluviometric variables without requiring long-term datasets. Partal and Kisis [23] assessed a wavelet analysis, in conjunction with neuro-fuzzy models, to forecast daily rainfall in Turkey. The developed models were significantly superior to traditional machine learning approaches, with a coefficient of determination (R^2) around 0.8–0.9. Li et al. [24] studied the performance of VMD, coupled with an extreme learning machine (ELM) model, to improve monthly rainfall forecasts in the northwest of China. This hybrid model highly outperformed traditional algorithms, with a meager computational cost, due to the non-training requirement of ELM. Filho and Lima [25] evaluated the singular spectrum analysis (SSA) forecasting monthly rainfall in Brazil. Based on the results, it could be concluded that the SSA caterpillar algorithm can deal with the inherent non-stationary nature of rainfall records, extracting its long varying trends and periodic components. Sun et al. [26] assessed SSA in Korea with linear recurrent formulas (LRF) and MLP. MLP obtained the best performance when forecasting monthly rainfall.

Finally, due to the significant advances in computation, deep learning algorithms are gaining very high popularity. In this sense, Kim et al. [27] evaluated the convolutional neural network (CNN), in conjunction with long short-term memory (LSTM), named convLSTM, to nowcast 1 and 2 h in advance, using two years dataset periods. ConvLSTM was able to reduce RMSE by 23%, when compared to linear regression. Ha et al. [28] developed a deep belief network model to forecast rainfall one day ahead in Seoul, performing better than MLP. Chen et al. [29] studied the performance of convLSTM with group normalization (GN) to improve the optimization process and employ a multi-sigmoid loss, inspired in the

critical success index (CSI) and compared to the COTREC model. COTREC obtained better performance, in terms of intensity in some areas, whereas convLSTM got a generally more reliable forecast.

This study aims to create a daily rainfall estimation model using only precipitation data, with different approaches in semiarid regions, such as Andalusia, to fill possible gaps in precipitation datasets. Additionally, a new approach is tested in daily rainfall estimations, which uses future precipitation values for this purpose. Thus, in this work, several machine learning models (MLP, SVM, and RF) and approaches for estimating missing rainfall data were tested and compared to empirical algorithms, such as linear interpolation (LI), in 14 locations from two different regions of Andalusia (coastal and inland areas) in Southern Spain. The first approach (A) uses neighbor stations' rainfall data of the same gap day and its distance to the target station. All these neighbor stations are located within a radius up to 50 km, following the recommendations of Barrios et al. [9] on a monthly basis and Estévez et al. [30] on a daily basis. Secondly, a new approach is considered, using only rainfall data (past and future values) from the target station as the model's inputs. Specifically, two different configurations were tested: (B) one day before and after the gap day and (C) two days before and after the gap day.

The rest of the work is organized as follows. Section 2 shows the information about the locations, the dataset, the theoretical background of the different machine learning (ML) models assessed, the preprocessing algorithms, and the evaluation metrics. Then, in Section 3, the results are reported and discussed. Finally, Section 4 describes the conclusions achieved in this work.

2. Materials and Methods

2.1. Source of Data

This study was carried out in Andalusia, Southern Spain, located in the southwest of Europe. Andalusia is a semiarid region with the following features: the meridians range from 1 to 7° W, the parallels from 37° to 39° N, an elevation above mean sea level from 26 to 822 m, and a total area of 87,268 m^2.

The datasets used belong to the Agroclimatic Information Network of Andalusia (RIAA) and can be downloaded at the following link: https://www.juntadeandalucia.es/agriculturaypesca/ifapa/riaweb/web (accessed on 30 July 2021). A total of 14 stations, divided into two areas (coastal and inland locations), were evaluated. The first group of areas included Jaen, La Higuera de Arjona, Linares, Mancha Real, Marmolejo, Sabiote, and Torreblascopedro and the second group included Málaga, Antequera, Archidona, Cártama, Churriana, Pizarra, and Vélez. Figure 1 shows their geographical locations, and Table 1 shows their geo-climatic characteristics.

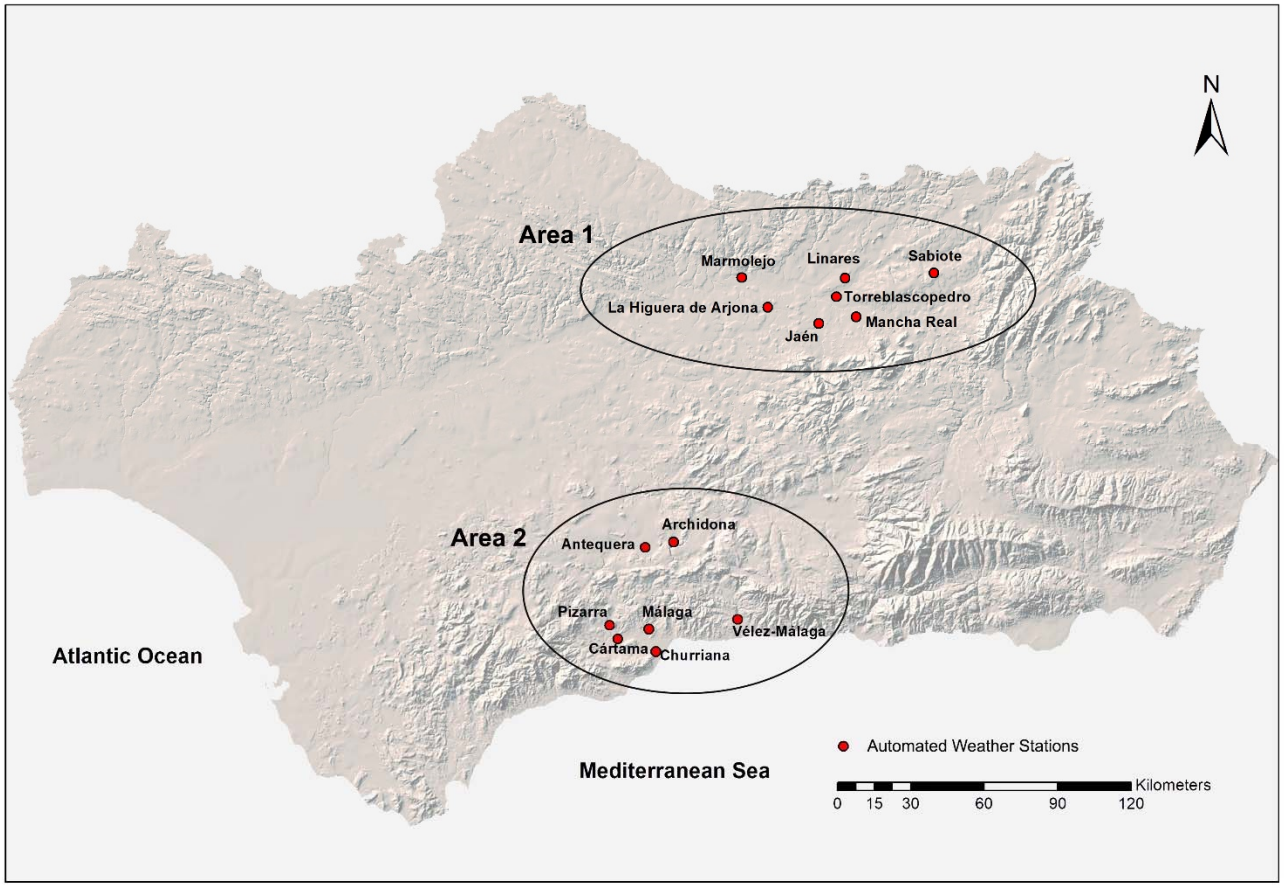

Figure 1. Spatial distribution of the fourteen automated weather stations used in this work.

Table 1. Geo-climatic characteristics of the AWS assessed in this work (lat.: latitude; long.: longitude; alt.: elevation above mean sea level).

Station	Alt. [m]	Lat. [°N]	Long. [°W]	Mean Annual Rainfall [mm]	Time-Period (Number of Days)
Area 1:					
Jaen (JAE)	299	37.89	3.77	446.54	From April 2001 to June 2021 (7361)
La Higuera de Arjona (ARJ)	257	37.95	4.00	477.68	From January 2001 to June 2021 (7456)
Linares (LIN)	432	38.07	3.65	466.70	From August 2000 to June 2021 (7601)
Mancha Real (MAN)	407	37.92	3.60	390.86	From August 2000 to June 2021 (7602)
Marmolejo (MAR)	208	38.06	4.13	523.36	From September 2000 to June 2021 (7590)
Sabiote (SAB)	791	38.08	3.24	446.98	From August 2000 to June 2021 (7615)
TorreblascoPedro (TOR)	275	37.99	3.69	434.37	From August 2000 to June 2021(7615)
Area 2:					
Antequera (ANT)	440	37.03	4.56	444.72	From November 2000 to June 2021 (7512)
Archidona (ARC)	516	37.08	4.43	457.83	From December 2000 to June 2021 (7483)
Cártama (CAR)	78	36.72	4.68	490.51	From June 2001 to June 2021 (7300)
Churriana (CHU)	17	36.67	4.50	510.32	From February 2001 to June 2021 (7426)
Málaga (MAL)	55	36.76	4.54	461.63	From October 2000 to June 2021 (7546)
Pizarra (PIZ)	71	36.77	4.72	463.47	From January 2001 to June 2021 (7447)
Vélez (VEL)	33	36.80	4.13	490.49	From October 2000 to June 2021 (7546)

2.2. Methodology

An essential prerequisite to guarantee reliable results using raw meteorological data is the application of quality assurance procedures. The quality control guidelines, reported by Estévez et al. [31], were followed, as well as the procedure to detect spurious precipitation signals from automated weather stations (AWS), also Estévez et al. [5].

Afterward, data preprocessing was required for every approach, obtaining the corresponding input configuration, according to every strategy (see Table 2). Three different methodologies were evaluated: approach (i)—the use of rainfall neighbor data and its distance data to estimate the precipitation values at a different site (all locations are located within a 50 km radius); approach (ii)—the use of one day before and ahead rainfall data values from the target station; and approach (iii)—the use of two days before and ahead rainfall data values from the target station.

Table 2. Inputs configurations of the different models and approaches assessed. DOY represents the day of year, P corresponds to precipitation, D corresponds to distance, and i represents an index to the dataset-specific day.

Target Station	Inputs Approach A	Inputs Approach B	Inputs Approach C
Area 1:			
Jaen	$DOY(i) + P_{ARJ}(i) + D_{JAE\text{-}LIN} + P_{LIN}(i) + D_{JAE\text{-}LIN} + P_{MAN}(i) + D_{JAE\text{-}MAN} + P_{MAR}(i) + D_{JAE\text{-}MAR} + P_{SAB}(i) + D_{JAE\text{-}SAB} + P_{TOR}(i) + D_{JAE\text{-}TOR}$	$DOY(i) + P_{JAE}(i-1) + P_{JAE}(i+1)$	$DOY(i) + P_{JAE}(i-1) + P_{JAE}(i-2) + P_{JAE}(i+1) + P_{JAE}(i+2)$
La Higuera de Arjona	$DOY(i) + P_{JAE}(i) + D_{ARJ\text{-}JAE} + P_{LIN}(i) + D_{ARJ\text{-}LIN} + P_{MAN}(i) + D_{ARJ\text{-}MAN} + P_{MAR}(i) + D_{ARJ\text{-}MAR} + P_{SAB}(i) + D_{ARJ\text{-}SAB} + P_{TOR}(i) + D_{ARJ\text{-}TOR}$	$DOY(i) + P_{ARJ}(i-1) + P_{ARJ}(i+1)$	$DOY(i) + P_{ARJ}(i-1) + P_{ARJ}(i-2) + P_{ARJ}(i+1) + P_{ARJ}(i+2)$
Linares	$DOY(i) + P_{JAE}(i) + D_{LIN\text{-}JAE} + P_{ARJ}(i) + D_{LIN\text{-}ARJ} + P_{MAN}(i) + D_{LIN\text{-}MAN} + P_{MAR}(i) + D_{LIN\text{-}MAR} + P_{SAB}(i) + D_{LIN\text{-}SAB} + P_{TOR}(i) + D_{LIN\text{-}TOR}$	$DOY(i) + P_{LIN}(i-1) + P_{LIN}(i+1)$	$DOY(i) + P_{LIN}(i-1) + P_{LIN}(i-2) + P_{LIN}(i+1) + P_{LIN}(i+2)$
Mancha Real	$DOY(i) + P_{JAE}(i) + D_{MAN\text{-}JAE} + P_{ARJ}(i) + D_{MAN\text{-}ARJ} + P_{LIN}(i) + D_{MAN\text{-}LIN} + P_{MAR}(i) + D_{MAN\text{-}MAR} + P_{SAB}(i) + D_{MAN\text{-}SAB} + P_{TOR}(i) + D_{MAN\text{-}TOR}$	$DOY(i) + P_{MAN}(i-1) + P_{MAN}(i+1)$	$DOY(i) + P_{MAN}(i-1) + P_{MAN}(i-2) + P_{MAN}(i+1) + P_{MAN}(i+2)$
Marmolejo	$DOY(i) + P_{JAE}(i) + D_{MAR\text{-}JAE} + P_{ARJ}(i) + D_{MAR\text{-}ARJ} + P_{LIN}(i) + D_{MAR\text{-}LIN} + P_{MAN}(i) + D_{MAR\text{-}MAN} + P_{SAB}(i) + D_{MAR\text{-}SAB} + P_{TOR}(i) + D_{MAR\text{-}TOR}$	$DOY(i) + P_{MAR}(i-1) + P_{MAR}(i+1)$	$DOY(i) + P_{MAR}(i-1) + P_{MAR}(i-2) + P_{MAR}(i+1) + P_{MAR}(i+2)$
Sabiote	$DOY(i) + P_{JAE}(i) + D_{SAB\text{-}JAE} + P_{ARJ}(i) + D_{SAB\text{-}ARJ} + P_{LIN}(i) + D_{SAB\text{-}LIN} + P_{MAN}(i) + D_{SAB\text{-}MAN} + P_{MAR}(i) + D_{SAB\text{-}MAR} + P_{TOR}(i) + D_{SAB\text{-}TOR}$	$DOY(i) + P_{SAB}(i-1) + P_{SAB}(i+1)$	$DOY(i) + P_{SAB}(i-1) + P_{SAB}(i-2) + P_{SAB}(i+1) + P_{SAB}(i+2)$
Torreblasco Pedro	$DOY(i) + P_{JAE}(i) + D_{TOR\text{-}JAE} + P_{ARJ}(i) + D_{TOR\text{-}ARJ} + P_{LIN}(i) + D_{TOR\text{-}LIN} + P_{MAN}(i) + D_{TOR\text{-}MAN} + P_{MAR}(i) + D_{TOR\text{-}MAR} + P_{SAB}(i) + D_{TOR\text{-}SAB}$	$DOY(i) + P_{TOR}(i-1) + P_{TOR}(i+1)$	$DOY(i) + P_{TOR}(i-1) + P_{TOR}(i-2) + P_{TOR}(i+1) + P_{TOR}(i+2)$
Area 2:			
Antequera	$DOY(i) + P_{ARC}(i) + D_{ANT\text{-}ARC} + P_{CAR}(i) + D_{ANT\text{-}CAR} + P_{CHU}(i) + D_{ANT\text{-}CHU} + P_{MAL}(i) + D_{ANT\text{-}MAL} + P_{PIZ}(i) + D_{ANT\text{-}PIZ} + P_{VEL}(i) + D_{ANT\text{-}VEL}$	$DOY(i) + P_{ANT}(i-1) + P_{ANT}(i+1)$	$DOY(i) + P_{ANT}(i-1) + P_{ANT}(i-2) + P_{ANT}(i+1) + P_{ANT}(i+2)$
Archidona	$DOY(i) + P_{ANT}(i) + D_{ARC\text{-}ANT} + P_{CAR}(i) + D_{ARC\text{-}CAR} + P_{CHU}(i) + D_{ARC\text{-}CHU} + P_{MAL}(i) + D_{ARC\text{-}MAL} + P_{PIZ}(i) + D_{ARC\text{-}PIZ} + P_{VEL}(i) + D_{ARC\text{-}VEL}$	$DOY(i) + P_{ARC}(i-1) + P_{ARC}(i+1)$	$DOY(i) + P_{ARC}(i-1) + P_{ARC}(i-2) + P_{ARC}(i+1) + P_{ARC}(i+2)$

Table 2. *Cont.*

Target Station	Inputs Approach A	Inputs Approach B	Inputs Approach C
Cártama	$DOY(i) + P_{ANT}(i) + D_{CAR\text{-}ANT} + P_{ARC}(i) + D_{CAR\text{-}ARC} + P_{CHU}(i) + D_{CAR\text{-}CHU} + P_{MAL}(i) + D_{CAR\text{-}MAL} + P_{PIZ}(i) + D_{CAR\text{-}PIZ} + P_{VEL}(i) + D_{CAR\text{-}VEL}$	$DOY(i) + P_{CAR}(i-1) + P_{CAR}(i+1)$	$DOY(i) + P_{CAR}(i-1) + P_{CAR}(i-2) + P_{CAR}(i+1) + P_{CAR}(i+2)$
Churriana	$DOY(i) + P_{ANT}(i) + D_{CHU\text{-}ANT} + P_{ARC}(i) + D_{CHU\text{-}ARC} + P_{CAR}(i) + D_{CHU\text{-}CAR} + P_{MAL}(i) + D_{CHU\text{-}MAL} + P_{PIZ}(i) + D_{CHU\text{-}PIZ} + P_{VEL}(i) + D_{CHU\text{-}VEL}$	$DOY(i) + P_{CHU}(i-1) + P_{CHU}(i+1)$	$DOY(i) + P_{CHU}(i-1) + P_{CHU}(i-2) + P_{CHU}(i+1) + P_{CHU}(i+2)$
Málaga	$DOY(i) + P_{ANT}(i) + D_{MAL\text{-}ANT} + P_{ARC}(i) + D_{MAL\text{-}ARC} + P_{CAR}(i) + D_{MAL\text{-}CAR} + P_{CHU}(i) + D_{MAL\text{-}CHU} + P_{PIZ}(i) + D_{MAL\text{-}PIZ} + P_{VEL}(i) + D_{MAL\text{-}VEL}$	$DOY(i) + P_{MAL}(i-1) + P_{MAL}(i+1)$	$DOY(i) + P_{MAL}(i-1) + P_{MAL}(i-2) + P_{MAL}(i+1) + P_{MAL}(i+2)$
Pizarra	$DOY(i) + P_{ANT}(i) + D_{PIZ\text{-}ANT} + P_{ARC}(i) + D_{PIZ\text{-}ARC} + P_{CAR}(i) + D_{PIZ\text{-}CAR} + P_{CHU}(i) + D_{PIZ\text{-}CHU} + P_{MAL}(i) + D_{PIZ\text{-}MAL} + P_{VEL}(i) + D_{PIZ\text{-}VEL}$	$DOY(i) + P_{PIZ}(i-1) + P_{PIZ}(i+1)$	$DOY(i) + P_{PIZ}(i-1) + P_{PIZ}(i-2) + P_{PIZ}(i+1) + P_{PIZ}(i+2)$
Vélez	$DOY(i) + P_{ANT}(i) + D_{VE\text{-}ANT} + P_{ARC}(i) + D_{VEL\text{-}ARC} + P_{CAR}(i) + D_{VEL\text{-}CAR} + P_{CHU}(i) + D_{VEL\text{-}CHU} + P_{MAL}(i) + D_{VEL\text{-}MAL} + P_{PIZ}(i) + D_{VEL\text{-}PIZ}$	$DOY(i) + P_{VEL}(i-1) + P_{VEL}(i+1)$	$DOY(i) + P_{VEL}(i-1) + P_{VEL}(i-2) + P_{VEL}(i+1) + P_{VEL}(i+2)$

Later, in order to tune all the different hyperparameters from the different models, train them, and evaluate their performance, the full dataset was split into training, validation, and test. The train (to fit all the final weights and biases from the final model) and test dataset (never-seen data to assess the performance) were randomly split into 70% and 30%, respectively. Prior to this stage, it is necessary to determine all the hyperparameters of the models (such as the number of hidden layers and neurons in a multilayer perceptron). To this purpose, the training dataset was divided into train_2 and validation (random 80% and 20%, respectively) to train and test the different hyperparameters until the fittest set is found. It is worth noting that the seed used in the random algorithm is the same in all cases, so all assessed models (from different approaches) have the same train, test, and validation dataset. Then, the Bayesian optimization algorithm took place, where different hyperparameters were tested, using the validation dataset, until the fittest set was found. Afterward, the entire train dataset from the initial split was used to adjust all the different weights and biases. Finally, the performance accuracy was assessed, using the testing dataset, which was never seen during previous processes. All this methodology is shown in a flowchart in Figure 2.

Besides, after splitting the dataset into train and test, a standardization was carried out, which is highly recommended to outperform machine learning models, especially neural network-based models [32]. This can be expressed as Equation (1):

$$x^* = \frac{x - \overline{x}}{\sigma} \quad (1)$$

where x represents the input data and $\overline{x}$ and σ correspond to the mean and standard deviation of the training dataset, respectively, and x^* is the standardized data.

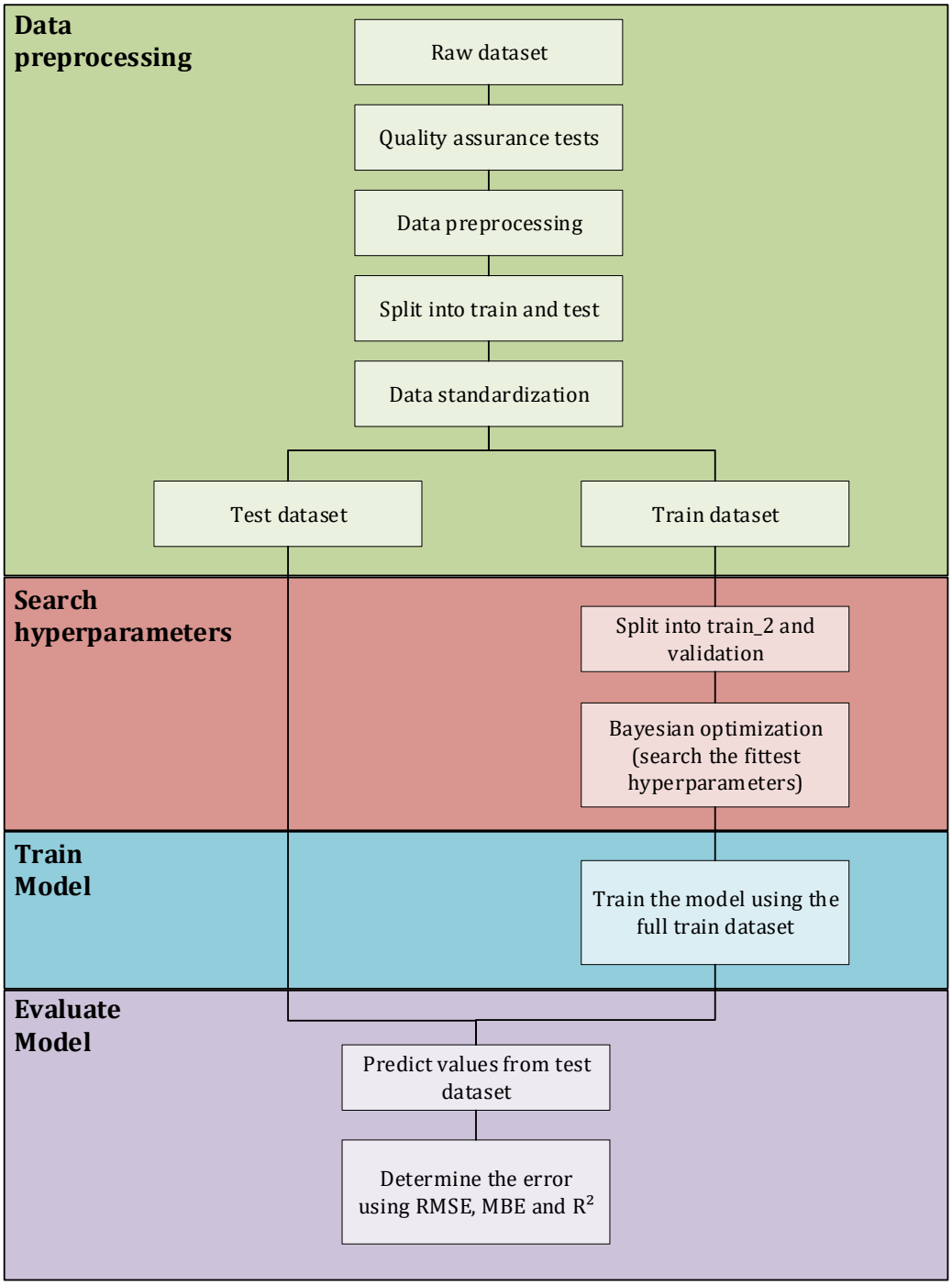

Figure 2. Methodology flowchart.

2.3. Multilayer Perceptron (MLP)

Multilayer perceptron is one of the most used models in different sectors, especially in hydrology [22,33]. Its functionality is based on neurons in the biological nervous system, where many interconnected neurons work together to generate an interaction, based on different stimuli. It is structured in three types of layers, the input and output correspond to the input and output of the model, respectively, as well as the hidden layer, where neurons are located. The activation function determines the output of a node, given a set of inputs. For example, rectified linear output (ReLU) represents a ramp for positive input values. The process in which the neurons learn (value adjustment of weights and biases) is carried out automatically, which is why this layer is called hidden. ADAM, a very common

algorithm for this purpose, uses squared gradients to scale the learning rate and a moving average of the gradients.

A single neuron mathematical logic is represented in Figure 3, where w represents the weight and b is the bias factor.

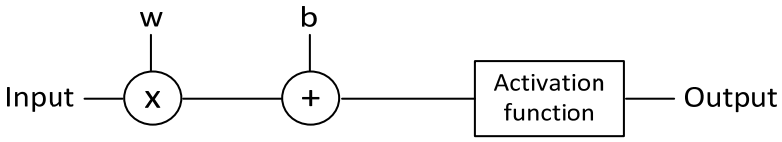

Figure 3. One neuron control logic.

For further information, the following works can be reviewed [34,35].

2.4. Support Vector Machine (SVM)

Support vector machine (SVM) is a supervised machine learning model that analyzes data for classification and regression tasks (also known as support vector regression (SVR)). For classification tasks, its functionality is based on searching the fittest hyperplane to separate different datapoints' classes (classification). On regression, it finds the hyperplane and margins that fit all of them (regression). Thus, an easy way to understand SVM for regression is similar to a linear regression, where a hyperplane (that includes the data) is searched, while having the flexibility to define how much error is considered acceptable. Figure 4 shows an example of SVM for classification (a) and regression (b).

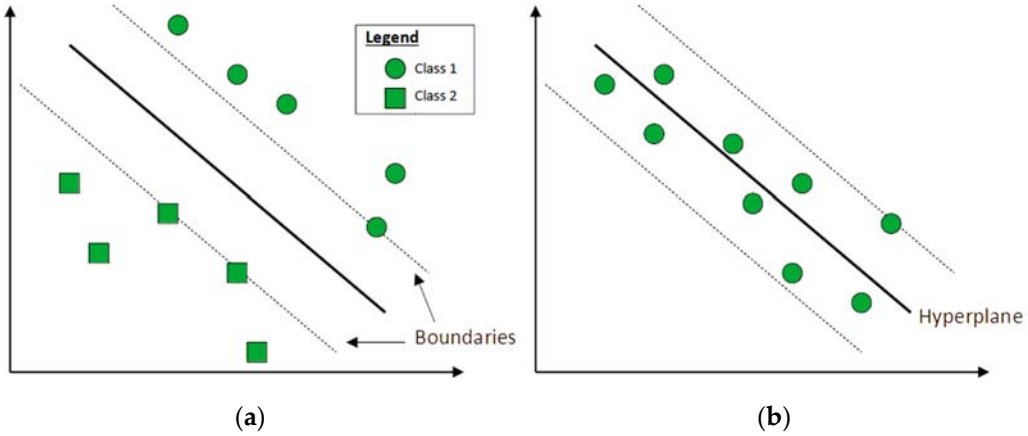

Figure 4. Support Vector Machine for classification (**a**) and regression (**b**).

The main feature of SVM models is the use of kernels (linear, sigmoid, or gaussian, among others) to enable operation in a high-dimensional feature map, where the number of features is greater than the number of observations.

SVM models are often used in rainfall forecasts, with promising results [36–38]. For further details, the following work can be reviewed [36,37].

2.5. Random Forest (RF)

Random Forest (RF) was first introduced by [39] as a supervised learning algorithm, where the "term" forest defines that it is built as an ensemble of decision tree models. The general idea is that the conjunction of multiple models increases the overall result. Additionally, RF introduces an extra-randomness when the number of trees starts to grow. Instead of searching for the best feature when splitting nodes, it searches for the best features among a random subset of them. The maximum number of features can be defined in scikit-learn as auto, sqrt, log2, none, or the exact number of maximum features (where auto and sqrt refer to the squared root of the initial number of features, log2 refers to the

logarithm base 2 of the number of features, and none is to use all the features). This results in a broader diversity and, as a consequence, a better final performance.

Other researchers have already assessed RF in rainfall with promising results [40–42]. For further details, the following work can be revised [42].

2.6. Bayesian Optimization

One of the critical aspects of machine learning models' efficiency is hyperparameter selection. Depending on whether the correct values have been set, the performance can dramatically change from outstanding to very poor results. A common practice in the scientific community uses a trial-and-error technique [22], where different values are evaluated, varying from dozens to thousands of possibilities. However, this method is far from efficient because if the hyperparameter space is ample, the algorithm (apart from being very slow) wastes significant time in non-promising configurations. On the other hand, when the hyperparameter space is tiny, an accurate hyperparameter configuration set may be missing, despite being quick. To solve this problem, several algorithms have been assessed in different works. In [6], the authors studied the effectiveness of particle-swarm optimization (PSO) and genetic algorithm (GA) to predict the monthly rainfall with MLP in a subtropical monsoon climate in Guilin, China. Wang et al. [43] assessed an artificial bee colony (ABC) with MLP to forecast rainfall values in 17 stations in the Wujiang River Basin. Banadkooki et al. [35] evaluated the flow regime optimization algorithm (FRA) with MLP and SVM to forecast monthly rainfall values in Iran.

In this study, Bayesian optimization was used, due to its high popularity in new automated machine learning (AML) models [44–47] and its good performance in [34,48]. It was first introduced by Wang et al. [43] as an algorithm, based on the Bayes theorem, to search the minimum/maximum function. Part of its popularity is due to its close relation to human behavior when tuning hyperparameters [49,50]. The prior results are taken into account to choose the following promising values to test, following the next four-step procedure: (1) the hyperparameter space is defined, which limits the values of the hyperparameter space; (2) the algorithm considers previous evaluations, in order to choose the following set of values to be assessed (acquisition function)—two kinds of possibilities can be handled, exploitation (consists of testing hyperparameters values that are assumed to be optimal) and exploration (the opposite of exploitation, to identify new best options); (3) to assess the new hyperparameter configuration using an objective function; and (4) if the optimization process has not finished yet, it goes to the second point. In this work, this algorithm was implemented using Python and the scikit-optimize library, following the instructions of Bellido-Jiménez et al. [34]. All the final hyperparameter sets, used for each model, approach, and location, can be seen in Table 3.

Table 3. Hyperparameter set for each model, approach, and location, after carrying out Bayesian optimization, where activation represents the activation function, the optimizer represents the optimizer function, epochs represents the number of epochs, neurons represents the number of hidden layers and the number of neurons of each, kernel is the kernel function, c and epsilon represent internal hyperparameters of SVR, n_estimators is the number of trees in RF, and max_features is the number of features to consider when looking for the best split.

Location	Models	Hyperparameters	Approaches: A	B	C
La Higuera de Arjona	MLP	activation	ReLU	ReLU	ReLU
		optimizer	ADAM	ADAM	ADAM
		epochs	100	87	53
		neurons	(20, 20)	(9, 15, 10)	(6, 15, 9)
	SVM	kernel	RBF	RBF	poly
		c	10.0	10.0	1.855
		epsilon	0.01	0.01	0.01
	RF	n_estimators	100	100	91
		max_features	sqrt	auto	log2

Table 3. *Cont.*

Location	Models	Hyperparameters	Approaches: A	B	C
Jaen	MLP	activation	ReLU	ReLU	ReLU
		optimizer	ADAM	ADAM	ADAM
		epochs	92	61	98
		neurons	(20, 20)	(2, 1, 12)	(1, 10, 8)
	SVM	kernel	linear	linear	RBF
		c	1.758	9.730	10.0
		epsilon	0.739	0.01	0.01
	RF	n_estimators	94	95	100
		max_features	auto	log2	log2
Linares	MLP	activation	ReLU	ReLU	ReLU
		optimizer	ADAM	ADAM	ADAM
		epochs	100	100	100
		neurons	(20, 20)	(1, 1, 1)	(1, 1, 1)
	SVM	kernel	linear	RBF	RBF
		c	10.0	4.023	10.0
		epsilon	0.01	0.018	0.01
	RF	n_estimators	100	97	80
		max_features	auto	sqrt	log2
Mancha Real	MLP	activation	ReLU	ReLU	ReLU
		optimizer	ADAM	ADAM	ADAM
		epochs	100	99	100
		neurons	(20, 20)	(5, 14)	(1, 1, 1)
	SVM	kernel	RBF	RBF	RBF
		c	10.0	6.235	9.211
		epsilon	0.01	0.010	0.01
	RF	n_estimators	75	41	46
		max_features	auto	sqrt	log2
Marmolejo	MLP	activation	ReLU	ReLU	ReLU
		optimizer	ADAM	ADAM	ADAM
		epochs	100	96	10
		neurons	(20, 6)	(5, 3, 11)	(1, 11)
	SVM	kernel	linear	RBF	RBF
		c	10.0	4.350	9.970
		epsilon	0.01	0.01	0.01
	RF	n_estimators	100	31	100
		max_features	auto	auto	sqrt
Sabiote	MLP	activation	ReLU	ReLU	ReLU
		optimizer	ADAM	ADAM	ADAM
		epochs	100	100	95
		neurons	(20, 20)	(1, 1, 1)	(2, 11, 9)
	SVM	kernel	linear	RBF	RBF
		c	10.0	10.0	10.0
		epsilon	0.01	0.01	0.01
	RF	n_estimators	72	39	57
		max_features	log2	log2	log2
Torreblascopedro	MLP	activation	ReLU	ReLU	ReLU
		optimizer	ADAM	ADAM	ADAM
		epochs	100	72	73
		neurons	(20, 12)	(1, 4, 13)	(5, 1, 17)
	SVM	kernel	linear	RBF	poly
		c	3.795	3.108	6.205
		epsilon	0.01	0.01	0.012
	RF	n_estimators	81	64	94
		max_features	log2	sqrt	sqrt

Table 3. *Cont.*

Location	Models	Hyperparameters	Approaches: A	B	C
Antequera	MLP	activation	ReLU	ReLU	ReLU
		optimizer	ADAM	ADAM	ADAM
		epochs	200	174	61
		neurons	(13, 8)	(5, 2, 20)	(14, 11, 13)
	SVM	kernel	linear	RBF	RBF
		c	8.684	7.627	4.981
		epsilon	0.225	0.01	0.014
	RF	n_estimators	55	94	41
		max_features	auto	auto	log2
Archidona	MLP	activation	ReLU	ReLU	ReLU
		optimizer	ADAM	ADAM	ADAM
		epochs	94	40	11
		neurons	(13, 5, 18)	(11, 12, 1)	(16, 11, 19)
	SVM	kernel	linear	poly	RBF
		c	7.246	4.531	4.104
		epsilon	0.01	0.01	0.013
	RF	n_estimators	81	93	100
		max_features	auto	auto	sqrt
Cártama	MLP	activation	ReLU	ReLU	ReLU
		optimizer	ADAM	ADAM	ADAM
		epochs	129	10	112
		neurons	(8, 13, 17)	(1, 1, 1)	(14, 17, 6)
	SVM	kernel	linear	RBF	poly
		c	6.273	7.830	3.862
		epsilon	0.01	0.01	0.01
	RF	n_estimators	92	10	38
		max_features	auto	sqrt	log2
Churriana	MLP	activation	ReLU	ReLU	ReLU
		optimizer	ADAM	ADAM	ADAM
		epochs	180	200	70
		neurons	(20, 20, 18)	(1, 1, 1)	(5, 15, 7)
	SVM	kernel	linear	RBF	RBF
		c	10.0	10.0	5.963
		epsilon	0.01	0.01	0.01
	RF	n_estimators	100	40	36
		max_features	log2	log2	sqrt
Málaga	MLP	activation	ReLU	ReLU	ReLU
		optimizer	ADAM	ADAM	ADAM
		epochs	158	97	127
		neurons	(20, 20, 20)	(17, 11, 10)	(13, 4, 16)
	SVM	kernel	linear	RBF	RBF
		c	8.784	9.999	6.952
		epsilon	0.01	0.01	0.011
	RF	n_estimators	69	14	10
		max_features	log2	log2	log2
Pizarra	MLP	activation	ReLU	ReLU	ReLU
		optimizer	ADAM	ADAM	ADAM
		epochs	192	94	171
		neurons	(13, 15, 8)	(3, 4, 6)	(14, 1, 6)
	SVM	kernel	linear	RBF	RBF
		c	7.642	10.0	4.031
		epsilon	0.015	0.01	0.01
	RF	n_estimators	76	45	95
		max_features	auto	sqrt	sqrt

Table 3. *Cont.*

Location	Models	Hyperparameters	Approaches: A	B	C
Vélez-Málaga	MLP	activation	ReLU	ReLU	ReLU
		optimizer	ADAM	ADAM	ADAM
		epochs	200	180	139
		neurons	(20, 20, 20)	(15, 13, 10)	(8, 2, 10)
	SVM	kernel	RBF	RBF	RBF
		c	10.0	6.032	10.0
		epsilon	0.01	0.01	0.01
	RF	n_estimators	72	62	78
		max_features	sqrt	log2	sqrt

2.7. Evaluation Metrics

To assess the efficiency of the developed models, the statistics root mean square error (RMSE), mean bias error (MBE), and coefficient of determination (R^2) were used. All of them are mathematically expressed as Equations (2)–(4), respectively:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(meas_i - pred_i)^2} \quad (2)$$

$$\mathrm{MBE} = \frac{1}{n}\sum_{i=1}^{n}(meas_i - pred_i) \quad (3)$$

$$R^2 = \frac{\left(\sum_{i=1}^{n}(meas_i - \mu_{meas})\left(pred_i - \mu_{pred}\right)\right)^2}{\sum_{i=1}^{n}(meas_i - \mu_{meas})^2 \sum_{i=1}^{n}\left(pred_i - \mu_{pred}\right)^2} \quad (4)$$

where n represents the number of prediction days, $meas_i$ corresponds to the measured value for a specific day, $pred_i$ is the predicted value, i represents every single gap day, and μ corresponds to the mean.

3. Results and Discussion

In order to help the reproducibility of this work, the best ML models were uploaded to an open access repository in Github (https://github.com/Smarity/gap-filling-precipitation-atmosphere-special-issue.git, accessed on 30 July 2021).

3.1. Using Neighbor Stations

Table 3 shows the RMSE, MBE, and R^2 performances for all locations in Area 1 (inland locations) using the first approach (A), information from other AWS located within 50 km. In Higuera de Arjona, MLP obtained the best RMSE and R^2 values (RMSE = 1.363 mm/day and R^2 = 0.894), very close to RF (RMSE = 1.384 mm/day and R^2 = 0.889). In terms of MBE, LI outperformed the rest of the ML models (MBE = −0.008 mm/day), followed closely to MLP and RF (MBE = 0.016 mm/day and MBE = 0.026 mm/day, respectively). In Jaen, all ML models outperformed LI in RMSE and R^2, where MLP obtained the best values (RMSE = 1.767 mm/day and R^2 = 0.827), whereas RF beat the rest, regarding MBE (MBE = 0.023 mm/day). In Linares, RF and LI obtained the best performance, in terms of MBE (MBE = 0.001 mm/day and MBE = −0.001 mm/day). Moreover, MLP outperformed the others, regarding RMSE and R^2 (RMSE = 1.723 mm/day and R^2 = 0.817), followed closely by RF (RMSE = 1.730 mm/day and R^2 = 0.815). In Mancha Real, MLP outperformed the other models in all statistics (RMSE = 1.662 mm/day, MBE = −0.072 mm/day, and R^2 = 0.831), whereas SVM was the worst (RMSE = 1.948 mm/day, MBE = −0.195 mm/day, and R^2 = 0.780). In Marmolejo, with the highest mean annual rainfall (523.36 mm/year),

the performance, in terms of RMSE and R^2, showed that RF obtained the best values (RMSE = 2.129 mm/day and R^2 = 0.801), followed closely by SVM (RMSE = 2.154 mm/day and R^2 = 0.795) and MLP (RMSE = 2.176 mm/day and R^2 = 0.791). In Sabiote, the location with the highest altitude, MLP obtained the best performance in RMSE and R^2 (RMSE = 2.049 mm/day and R^2 = 0.752), but LI beat ML in MBE (MBE = −0.006 mm/day). Finally, in Torreblascopedro, SVM outperformed the rest for all statistics (RMSE = 1.246 mm/day, MBE = −0.005, and R^2 = 0.894), being the most accurate from this first region. It is worth noting that MLP generally obtained the best results, regarding RMSE and R^2, in most locations, whereas RF and LI obtained the best values for MBE. Additionally, even though ML outperformed LI in all locations, the average improvement was not very significant.

Tables 4 and 5 shows the RMSE, MBE, and R^2 values for all locations and models in the coastal locations (Area 2). In Antequera, MLP beat the other models for all statistics (RMSE = 1.596 mm/day, MBE = 0.035 mm/day, and R^2 = 0.875), sharing the same R^2 performance with SVM (R^2 = 0.875). All ML models highly outperformed LI, considering all statistics (especially RMSE and R^2), except for MBE using SVM. In Archidona, MLP also obtained the most accurate modeling in RMSE and R^2 (RMSE = 1.811 mm/day and R^2 = 0.844), followed closely to SVM (RMSE = 1.817 mm/day and R^2 = 0.844). Regarding MBE, RF outperformed the rest (MBE = −0.019 mm/day). In Cártama, RF obtained the best MBE value (MBE = 0.002 mm/day), whereas SVM got the best RMSE and R^2 performance (RMSE = 2.502 mm/day and R^2 = 0.778). In Churriana, MLP highly outperformed the rest, in terms of RMSE and R^2 (RMSE = 2.192 mm/day and R^2 = 0.876), whereas RF beat MLP in MBE (MBE = 0.019 mm/day and MBE = −0.052 mm/day, respectively). In Málaga, RF obtained the best values for RMSE and MBE (RMSE = 2.433 mm/day and MBE = 0.012 mm/day), whereas MLP got the most accurate values for R^2 (R^2 = 0.830). In Pizarra, all models obtained very similar performance (even LI). RMSE ranged from 2.032 mm/day (by MLP and SVM) to 2.108 mm/day (by LI), MBE ranged from 0.039 mm/day (by RF) to −0.112 mm/day (by SVM), and R^2 ranged from 0.842 (by LI) to 0.854 (by MLP). Finally, in Vélez, MLP outperformed the rest of the models, in terms of RMSE and R^2 (RMSE = 3.219 mm/day and R^2 = 0.742), while RF obtained the best MBE performance (MBE = −0.020 mm/day), followed closely to MLP (MBE = −0.074 mm/day). Generally, the results obtained by ML highly outperformed LI in most locations and statistics, except for MBE, in which LI obtained very accurate results. Thus, the use of ML models to gap-fill daily rainfall data is highly recommended for coastal sites, performing significantly better than LI, arising the effect of sea distance in rainfall modelling. Eventually, in Figure 5, all these RMSE, MBE, and R^2 values, from both areas and all models, are represented in a scatter plot. Due to the different performances between the ML models, it can be stated that MLP obtained the best results, or very close to them, in most locations. On the other hand, SVM had accurate performances in coastal sites, whereas the behavior was not so good on inland locations. Finally, RF behaved opposite to SVM, having an accurate performance on inland locations and a worse modeling on inland sites.

Table 4. RMSE, MBE, and R^2 performance values from testing dataset for all locations and models in the first area (inland locations), using data from neighbor stations. The best values for each site are in bold.

Stations (Area 1)	Model	RMSE [mm/day]	MBE [mm/day]	R^2
La Higuera de Arjona	MLP	**1.363**	0.016	**0.894**
	SVM	1.800	−0.106	0.818
	RF	1.384	0.026	0.889
	LI	1.502	**−0.008**	0.869

Table 4. *Cont.*

Stations (Area 1)	Model	RMSE [mm/day]	MBE [mm/day]	R^2
Jaen	MLP	**1.767**	−0.097	**0.827**
	SVM	1.822	−0.064	0.823
	RF	1.880	**0.023**	0.804
	LI	1.916	0.051	0.797
Linares	MLP	**1.723**	0.083	**0.817**
	SVM	1.808	−0.106	0.798
	RF	1.730	**0.001**	0.815
	LI	1.896	**−0.001**	0.784
Mancha Real	MLP	**1.662**	**−0.072**	**0.831**
	SVM	1.948	−0.195	0.780
	RF	1.730	−0.078	0.816
	LI	1.852	0.110	0.790
Marmolejo	MLP	2.176	−0.187	0.791
	SVM	2.154	−0.169	0.795
	RF	**2.129**	**0.041**	**0.801**
	LI	2.392	−0.249	0.753
Sabiote	MLP	**2.049**	−0.101	**0.752**
	SVM	2.135	−0.224	0.739
	RF	2.105	−0.061	0.740
	LI	2.112	**−0.006**	0.742
Torreblascopedro	MLP	1.270	−0.035	0.894
	SVM	**1.246**	**−0.005**	**0.898**
	RF	1.359	0.019	0.878
	LI	1.277	0.047	0.894
Mean values		1.792	−0.048	0.815

Table 5. RMSE, MBE, and R^2 performance values from testing dataset for all locations and models in the second area (coastal locations), using data from neighbor stations. The best values for each site are in bold.

Stations (Area 2)	Model	RMSE [mm/day]	MBE [mm/day]	R^2
Antequera	MLP	**1.595**	**0.035**	**0.875**
	SVM	1.632	−0.104	**0.875**
	RF	2.009	0.042	0.799
	LI	2.839	0.100	0.684
Archidona	MLP	**1.811**	−0.043	**0.844**
	SVM	1.817	−0.168	**0.844**
	RF	2.002	**−0.019**	0.809
	LI	3.286	−0.041	0.594
Cártama	MLP	2.640	−0.075	0.756
	SVM	**2.502**	−0.106	**0.778**
	RF	2.820	**0.002**	0.737
	LI	2.630	0.061	0.756
Churriana	MLP	**2.192**	−0.052	**0.876**
	SVM	2.465	−0.147	0.860
	RF	2.315	**0.019**	0.862
	LI	2.973	−0.061	0.790
Málaga	MLP	2.485	0.099	**0.830**
	SVM	2.448	−0.170	0.825
	RF	**2.433**	**0.012**	0.816
	LI	2.610	0.04	0.785

Table 5. *Cont.*

Stations (Area 2)	Model	RMSE [mm/day]	MBE [mm/day]	R^2
Pizarra	MLP	**2.032**	0.043	**0.854**
	SVM	2.083	−0.112	0.853
	RF	**2.032**	**0.039**	**0.854**
	LI	2.108	0.079	0.842
Vélez-Málaga	MLP	**3.219**	−0.074	**0.742**
	SVM	3.531	−0.376	0.706
	RF	3.306	**−0.020**	0.719
	LI	3.489	−0.157	0.692
Mean values		2.475	−0.041	0.794

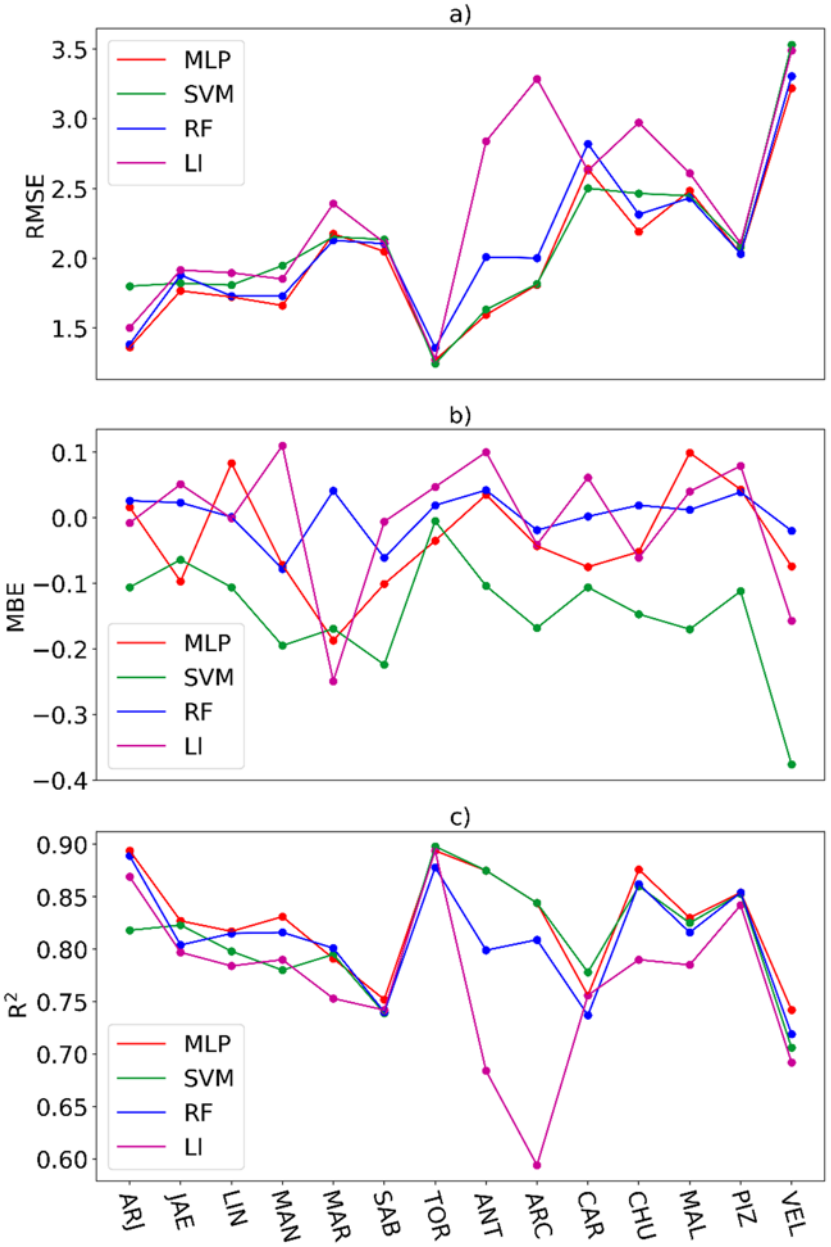

Figure 5. RMSE (**a**), MBE (**b**), and R^2 (**c**) values from testing dataset for all stations and models (MLP, SVM, RF, and LI), using only precipitation data from neighbor stations.

3.2. Using Data from the Target Station

Tables 6 and 7 show the RMSE and MBE values for the inland and coastal locations, using two different approaches, one day after and before (approach B) and two days after and before (approach C), as inputs, respectively. Generally, all the results are much worse than in Tables 2 and 3, for all cases. Mancha Real obtained an RMSE value above 4.0 for all models, whereas Churriana got the worse values (RMSE > 6.0 mm/day). In terms of MBE, La Higuera de Arjona obtained the best performance (MBE = −0.021 mm/day) using MLP and approach B, whereas Marmolejo was the worst (MBE = −1.459 mm/day), using MLP and this same approach. Finally, in terms of R^2, the values obtained are low, from R^2 = 0.004 (by SVM in Archidona) to R^2 = 0.079 (by MLP in Málaga), highlighting the non-autocorrelation between precipitation values from the previous and following days. Comparing the results between B and C, in terms of RMSE, on average, approach C (RMSE = 4.359 mm/day) obtained a slightly better performance than approach B (RMSE = 4.323 mm/day). However, in area 2, the use of approach B (RMSE = 4.986 mm/day) was significantly better than approach C (RMSE = 5.588 mm/day).

Table 6. RMSE, MBE, and R^2 performance values from testing dataset for all locations and models in the first area (inland locations), using data from the target station in two different approaches, with the use of the previous and following day and the use of the two previous and two following days. The best values from each station are in bold.

Stations (Area 1)	Model	One Day (B) RMSE [mm/day]	One Day (B) MBE [mm/day]	One Day (B) R^2	Two Days (C) RMSE [mm/day]	Two Days (C) MBE [mm/day]	Two Days (C) R^2
La Higuera de Arjona	MLP	**4.409**	**−0.021**	**0.023**	**4.079**	**0.061**	**0.051**
	SVM	4.601	−1.218	0.008	4.348	−1.225	0.027
	RF	4.524	−0.880	0.020	4.224	−0.932	0.033
Jaen	MLP	3.875	−1.071	0.016	**4.423**	**−0.016**	**0.022**
	SVM	3.857	−1.039	0.018	4.613	−1.189	0.007
	RF	**3.785**	**−0.771**	**0.019**	4.583	−1.103	0.011
Linares	MLP	4.797	−1.378	0.015	4.455	−1.260	0.010
	SVM	4.754	−1.308	**0.019**	4.423	−1.202	0.010
	RF	**4.719**	**−0.940**	0.012	**4.371**	**−0.911**	**0.014**
Mancha Real	MLP	**3.246**	**0.128**	**0.047**	**3.288**	**0.305**	**0.012**
	SVM	3.450	−0.946	0.005	3.390	−0.842	0.002
	RF	3.386	−0.820	0.021	3.383	−0.788	0.003
Marmolejo	MLP	5.530	−1.459	0.012	5.396	−1.374	0.014
	SVM	5.501	−1.410	0.022	5.360	−1.307	0.015
	RF	**5.474**	**−0.947**	**0.014**	**5.235**	**−0.761**	**0.028**
Sabiote	MLP	3.992	−1.159	0.026	4.186	−1.114	0.008
	SVM	3.937	−1.091	0.030	4.155	−1.041	0.006
	RF	**3.893**	**−0.797**	**0.016**	**4.119**	**−0.910**	**0.010**
Torreblascopedro	MLP	4.658	−1.287	0.022	4.283	−1.204	0.011
	SVM	4.626	−1.236	0.027	4.263	−1.167	0.010
	RF	**4.539**	**−0.900**	**0.021**	**4.202**	**−0.802**	**0.015**
Mean values		4.359	−0.978	0.034	4.322	−0.894	0.037

Table 7. RMSE, MBE, and R^2 performance values from testing dataset for all locations and models in the second area (coastal locations), using data from the target station in two different approaches, with the use of the previous and following day and the use of the two previous and two following days. The best values from each station are in bold.

Stations (Area 2)	Model	One Day (B) RMSE [mm/day]	One Day (B) MBE [mm/day]	One Day (B) R^2	Two Days (C) RMSE [mm/day]	Two Days (C) MBE [mm/day]	Two Days (C) R^2
Antequera	MLP	**5.035**	**−0.246**	**0.027**	4.521	−1.246	**0.048**
	SVM	5.243	−1.296	0.021	4.480	−1.197	0.045
	RF	5.229	−1.221	0.005	**4.467**	**−1.163**	0.017
Archidona	MLP	4.108	−1.095	0.008	**4.109**	**−0.059**	**0.041**
	SVM	4.089	−1.012	0.004	4.328	−1.180	0.027
	RF	**4.083**	**−0.480**	**0.023**	4.252	−0.695	0.029
Cártama	MLP	**5.479**	−1.149	0.009	**5.235**	**0.239**	**0.040**
	SVM	5.550	−1.144	**0.027**	5.431	−1.132	0.021
	RF	5.631	**−0.896**	0.021	5.374	−1.054	0.024
Churriana	MLP	6.551	−1.314	**0.051**	6.849	−1.406	**0.017**
	SVM	6.449	−1.263	0.045	6.817	−1.367	0.009
	RF	**6.448**	**−1.148**	0.022	**6.781**	**−1.263**	0.012
Málaga	MLP	**5.028**	**0.294**	**0.079**	6.850	−1.324	0.044
	SVM	5.279	−1.028	0.023	6.765	−1.273	**0.056**
	RF	5.104	−0.884	**0.079**	**6.693**	**−1.182**	0.041
Pizarra	MLP	**5.152**	**0.253**	0.031	**5.871**	**0.014**	0.050
	SVM	5.266	−1.071	**0.044**	6.064	−1.205	**0.081**
	RF	5.267	−0.785	0.025	6.058	−1.074	0.021
Vélez-Málaga	MLP	**5.295**	**0.191**	**0.076**	**5.360**	**0.149**	**0.061**
	SVM	5.535	−1.198	0.047	5.544	−1.144	0.040
	RF	5.465	−1.046	0.052	5.489	−1.054	0.056
Mean values		5.299	−0.835	0.019	5.587	−0.934	0.015

Finally, in Figures 5 and 6, all the RMSE, MBE, and R^2 values, from both areas and all models, are represented in a scatter plot.

3.3. Comparison of the Two Areas

In order to compare the results in the two different areas, Figure 7 shows the RMSE, MBE, and R^2 performance values for these two kinds of locations (inland and coastal), using the best approach (data from neighbor stations). In terms of RMSE mean values, MLP outperformed RF and SVM. Besides, the models applied on the coastal locations underperformed, on average, in all cases and obtained higher variability across sites, rather than inland ones. In terms of MBE mean values, RF and LI obtained values very close to 0, whereas SVM overestimated in most stations. Finally, in terms of R^2, the results by ML models were quite similar in both inland and coastal locations. However, the results of LI were significantly worse than ML in coastal sites, whereas SVM performed worse on inland sites than coastal.

Additionally, Table 8 displays the maximum improvement, in terms of RMSE, R^2, and MBE, comparing ML to LI (using the first approach). In inland sites, the RMSE improvement ranged from 0.031 mm/day in Torreblascopedro to 0.263 mm/day in Marmolejo, as well as from 0.004 (Torreblascopedro) to 0.048 (Marmolejo), in terms of R^2. On the other hand, the upgrades in coastal sites ranged from RMSE = 0.076 mm/day and R^2 = 0.012 (in Pizarra) to RMSE = 1.475 mm/day and R^2 = 0.25 (in Archidona). Thus, coastal locations significantly differed between linear interpolation and ML models for gap-filling daily rainfall. In contrast, in inland areas, the improvement was not substantial.

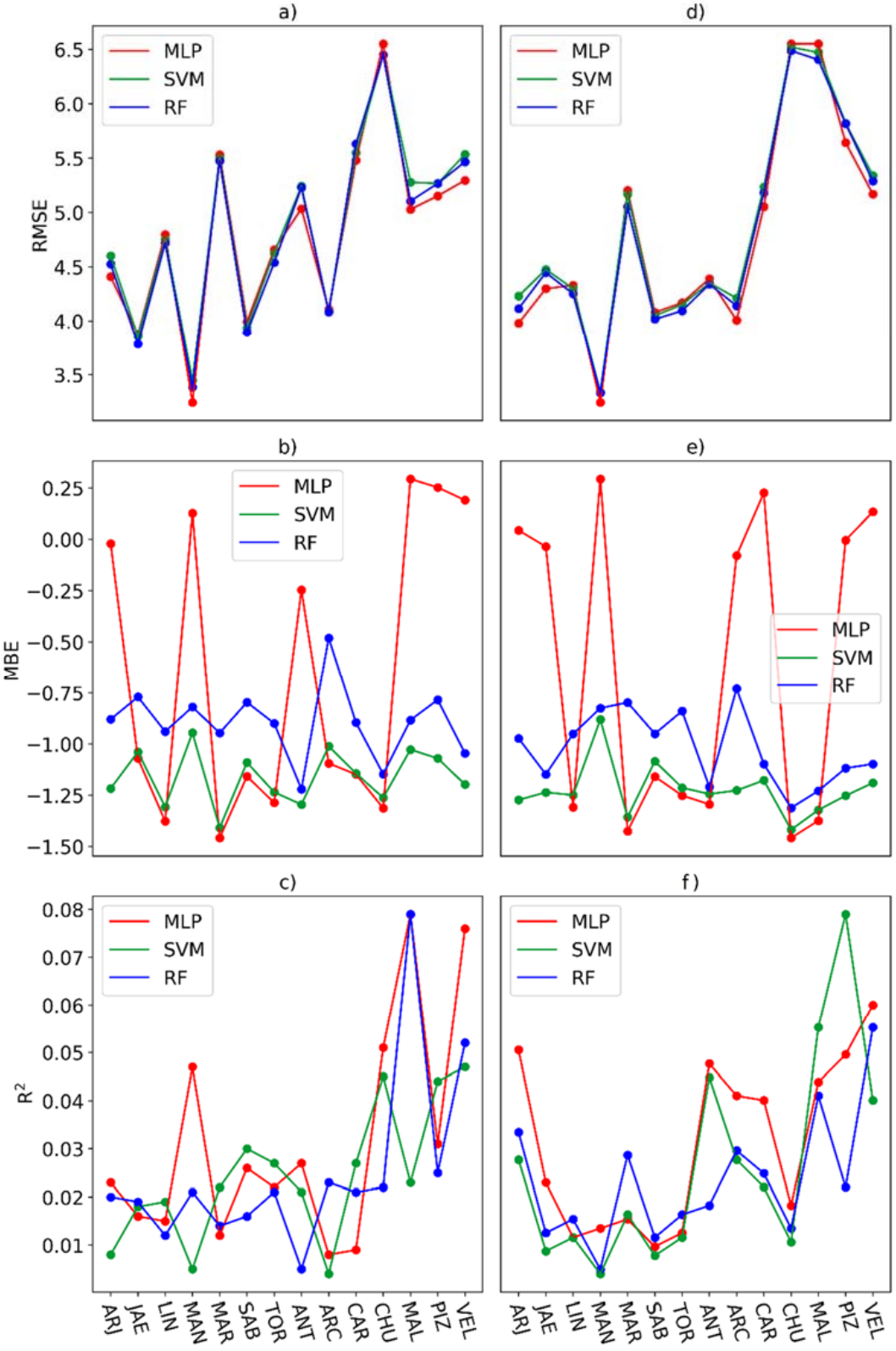

Figure 6. RMSE, MBE, and R^2 values from testing dataset for all stations and models (MLP, SVM, and RF), using only precipitation data from the target station. (**a**) RMSE using approach B, (**b**) MBE using approach B, (**c**) R^2 using approach B, (**d**) RMSE using approach C, (**e**) MBE using approach C, (**f**) R^2 using approach C.

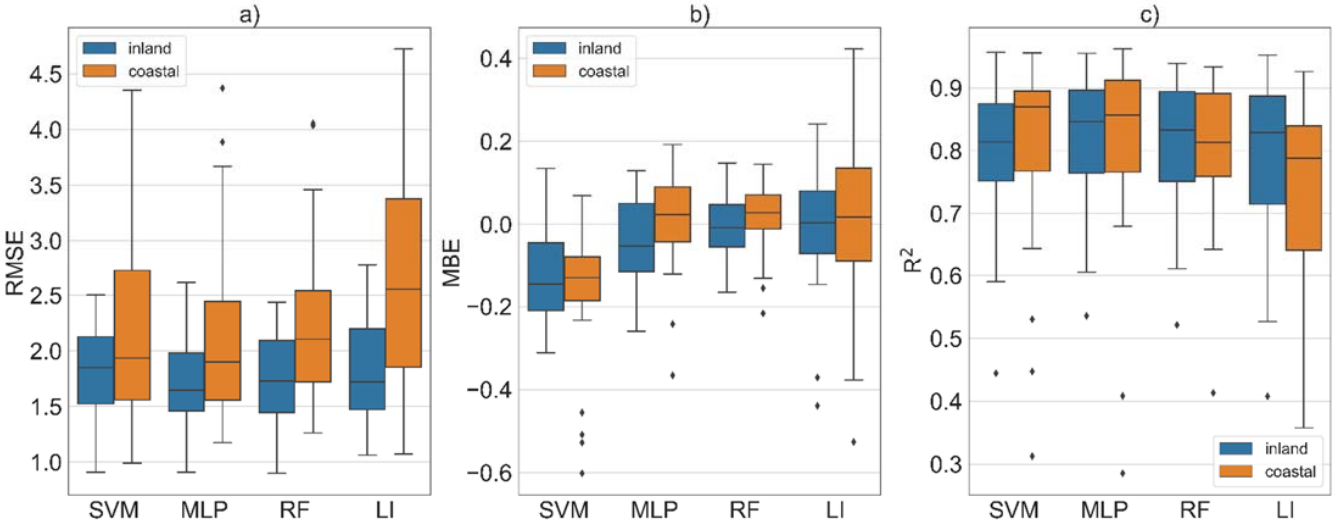

Figure 7. RMSE (**a**), MBE (**b**), and R^2 (**c**) values of the different models (SVM, MLP, and RF) in the coastal and inland locations, using rainfall values from neighbor stations as inputs, where the minimum, the first interquartile (Q1), the median, the third interquartile (Q3), the maximum, and the outlier values are represented.

Table 8. Best improvements between simple arithmetic averaging and the best ML model from each site for R^2 and RMSE. A positive value means that ML outperformed LI.

Station	RMSE (mm/day)	R^2
La Higuera de Arjona	0.139	0.025
Jaén	0.149	0.03
Linares	0.173	0.033
Mancha Real	0.19	0.041
Marmolejo	0.263	0.048
Sabiote	0.063	0.010
Torreblascopedro	0.031	0.004
Antequera	1.244	0.191
Archidona	1.475	0.25
Cártama	0.128	0.022
Churriana	0.781	0.086
Málaga	0.177	0.045
Pizarra	0.076	0.012
Vélez-Málaga	0.265	0.05

Thus, using empirical approaches (such as LI) to gap-fill daily rainfall data is not recommended, especially in coastal sites; the results are worse than ML, due to the effect of sea distance.

3.4. Seasonality Performance

In order to assess seasonal performance, the RMSE, MBE, and R^2 of all the stations and approaches, for the different evaluated models (SVM, MLP, and RF), are represented in Figure 8. Regarding RMSE, summer, autumn, and spring obtained very similar average performances, whereas, in winter, the mean results were the worst. Moreover, summer obtained the narrowest interquartile range, but spring and winter got the more extensive range, with LI being the model with the worst range (the less confident between different stations) among all seasons and models. MBE, MLP, RF, and LI always performed with very similar average results, although LI had the widest interquartile range for all seasons. Besides, SVM always performed the worst, in terms of MBE. In terms of R^2, the highest mean values were carried out in winter, whereas the worst results were achieved in summer and spring. Regarding mean, all models performed with similar values during the same season.

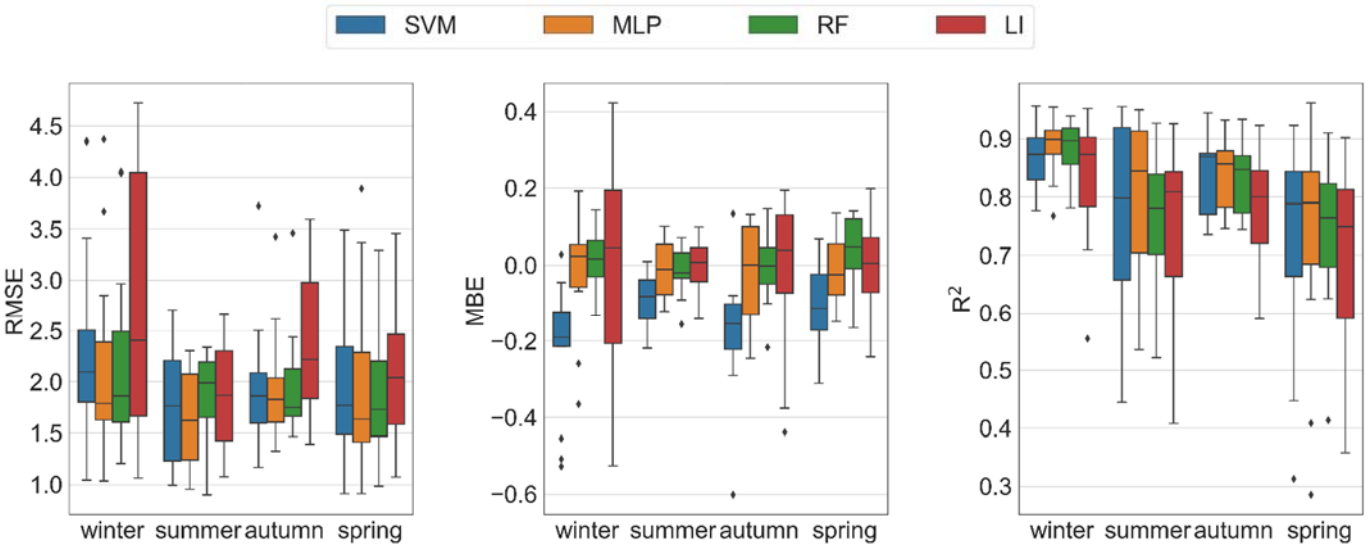

Figure 8. Seasonality performance of the different models (SVM, MLP, and RF) in all the stations and approaches.

Additionally, in Figures 9 and 10, the values predicted by the different ML models using the first approach are shown and compared to LI. In Figure 9, the predictions from Torreblascopedro are plotted (the site that obtained the best performance, in terms of RMSE and R^2). In winter, all predictions are close to the 1:1 line, which denotes the excellent performance of this model during this season. The predictions were also close to the 1:1 line in spring and autumn, although the points were more dispersed than in winter. Finally, summer obtained the worst results, with farthest points to the 1:1 line, especially with high rainfall values.

Finally, Figure 10 plots the prediction rainfall values in Archidona. Spring obtained the best general predictions among all models, followed by autumn, summer, and winter, in this order. The highest differences between ML and LI were found in winter and autumn, where most LI predictions were farther from the 1:1 line.

Generally, summer obtained worse results than the rest of the seasons, due to the Mediterranean climate; during summers in Andalusia, precipitation is very occasional. They usually respond to local events, such as local torments. So, gap-filling rainfall data using neighbor stations with very different pluviometry makes models fail in those specific dates. Comparing the results between Torreblascopedro and Archidona, the most significant differences can be seen in winter, where LI performed much worse than ML approaches.

3.5. General Discussion

In terms of R^2, the results obtained in this work outperformed those obtained by Kim and Ryu [51] (Pocatello, ID, USA) using IDWM, OK, and GME, in conjunction with cluster analysis, having the best R^2 performance, with a value below 0.7 (R^2 = 0.689 or R = 0.83). Besides, the models developed in this work highly improved the RMSE and R^2 performance of Wuthiwongyothin [52] in Northern Thailand, using the K-means technique with the inverse distance weighting (IDW) and correlation coefficient weighting (CCW), where the mean R^2 values among all stations were below 0.6. Moreover, in terms of R^2, the values obtained by Sehad et al. [53] in North Algeria using multispectral MSG SEVIRI imagery were slightly worse, on average, than the obtained in this work, with a mean R^2 = 0.7241. However, in absolute terms, its developed model outperformed this work's best results (R^2 = 0.921 against R^2 = 0.898 in Torreblascopedro using SVM). Thus, ML models with neighbor station data located within a 50 km radius are highly recommended to gap-fill rainfall values in coastal locations, due to their accurate performance (among other approaches) in the different areas assessed along the Andalusia region, being the preferred use of neighbor stations, over the use of cluster analysis with stations located

within a further radius distance. However, in inland sites, the performances carried out by ML against LI were not as significant as in coastal sites. Finally, in order to improve the state of the art of these approaches, future works could analyze the possibility of false alarms and missing rainfall cases using the models developed in this work.

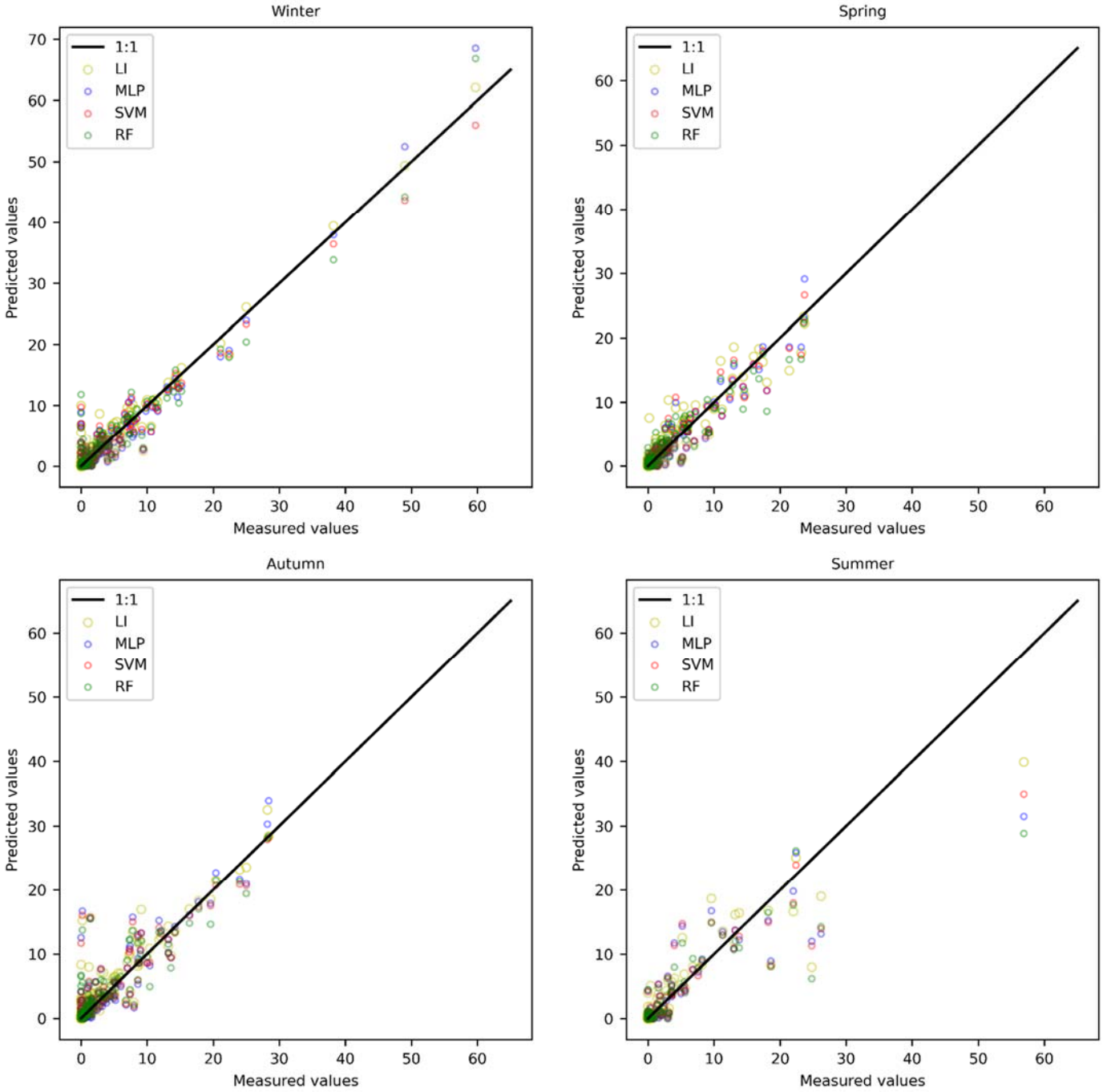

Figure 9. Scatter plot for predicted values in Torreblascopedro, using MLP, SVM, RF, and LI, during the different seasons.

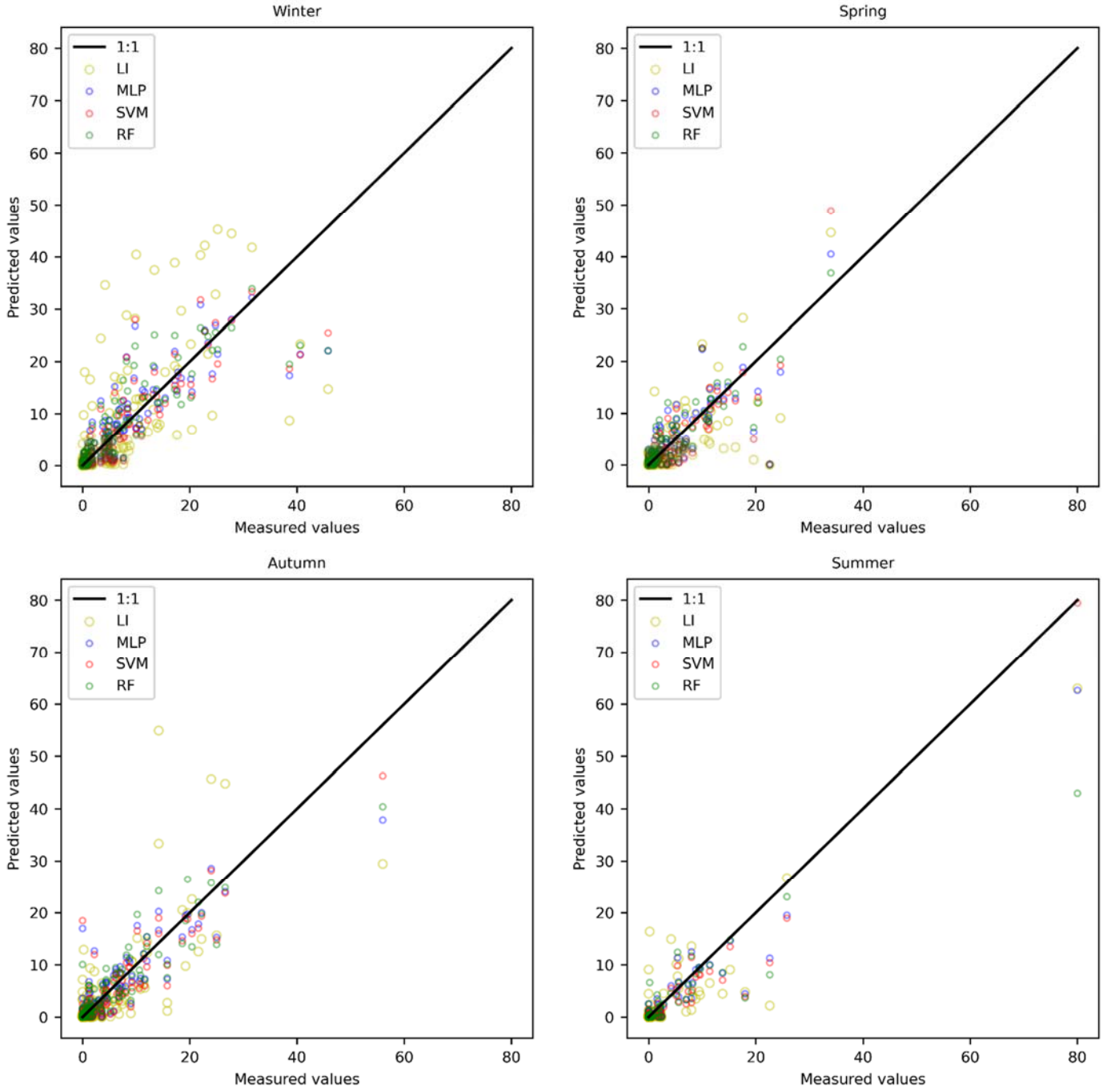

Figure 10. Scatter plot for predicted values in Archidona, using MLP, SVM, RF, and LI, during the different seasons.

4. Conclusions

Three different approaches were evaluated for gap-filling daily rainfall values: (A) the use of data from neighbor stations within 50 km, (B) the use of one day before and ahead from the target station, and (C) the use of two days before and ahead from the target station. Fourteen different locations were evaluated from two areas, corresponding to inland and coastal sites. Additionally, three different ML models were assessed for this purpose: MLP, SVM, and RF. Daily large datasets of around 21 years were used (from 2000 to 2021), where 70% was used for training and a random 30% for testing purposes. Besides, 20% from the training dataset was used to find the fittest hyperparameters. Finally, a seasonality analysis was carried out. Based on the arisen results, no ML model significantly outperformed the rest, although MLP obtained the best results, or very close to them, in most locations. On

the other hand, SVM had accurate performances in coastal sites, whereas the behavior was not so good at inland locations. RF behaved the opposite to SVM, having an accurate performance at inland locations and worse modeling at inland sites. Moreover, the first approach (the use of neighbor data) was notably better than the other approaches, with RMSE values below 2.0 mm/day and R^2 values above 0.85 in most stations. There were no significant seasonal differences in performance, in terms of RMSE and MBE values in winter, spring, and autumn, but the results obtained in summer were generally worse for all locations. Besides, coastal area location models performed slightly worse and with higher performance differences between ML and LI, in most sites and models, highlighting the differences in rainfall prediction efficiency, depending on the sea distance. In conclusion, it could be stated that the use of neighbor data with MLP is highly recommended as a rainfall gap-filling technique, rather than the use of data from the target station from the past and future. Moreover, when these work's results are compared to different paper's approaches using a cluster analysis from wider ranges, the use of closer stations (within a 50 km radius) obtained better results in terms of R^2.

Finally, due to the significant need to have a complete time series rainfall dataset on a daily basis and the increasing interest in installing low-cost wireless sensors (IoT), the models developed and assessed in this work can help with gap-filling datasets in this work near-real-time, thanks to the decreasing price of the low-cost automated weather stations using these new devices.

Author Contributions: Conceptualization, A.P.G.-M., J.A.B.-J. and J.E.G.; methodology, J.A.B.-J. and J.E.G.; software, J.A.B.-J. and J.E.G.; validation, A.P.G.-M., J.A.B.-J. and J.E.G.; formal analysis, J.A.B.-J. and J.E.G.; investigation, J.A.B.-J. and J.E.G.; resources, A.P.G.-M., J.A.B.-J. and J.E.G.; data curation, J.A.B.-J. and J.E.G.; writing—original draft preparation, J.A.B.-J. and J.E.G.; writing—review and editing, A.P.G.-M., J.A.B.-J. and J.E.G.; visualization, J.A.B.-J. and J.E.G.; supervision, A.P.G.-M., J.A.B.-J. and J.E.G.; project administration, A.P.G.-M. and J.E.G.; funding acquisition, A.P.G.-M. and J.E.G. All authors have read and agreed to the published version of the manuscript.

Funding: Spanish Ministry of Science, Innovation and Universities [grant AGL2017-87658-R] and University of Cordoba: PIF scholarship.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. These data can be found here: https://www.juntadeandalucia.es/agriculturaypesca/ifapa/riaweb/web (accessed on 30 July 2021).

Acknowledgments: J.A. Bellido-Jiménez wishes to thank the University of Córdoba for providing a PIF scholarship funded by the research program and the Spanish Ministry of Science, Innovation and Universities. Grant number AGL2017-87658-R for also funding this research.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Shen, R.; Huang, A.; Li, B.; Guo, J. Construction of a drought monitoring model using deep learning based on multi-source remote sensing data. *Int. J. Appl. Earth Obs. Geoinf.* **2019**, *79*, 48–57. [CrossRef]
2. Fernández, A.J.; Molero, F.; Becerril-Valle, M.; Coz, E.; Salvador, P.; Artíñano, B.; Pujadas, M. Application of remote sensing techniques to study aerosol water vapour uptake in a real atmosphere. *Atmos. Res.* **2018**, *202*, 112–127. [CrossRef]
3. Astel, A.; Mazerski, J.; Polkowska, Z.; Namieśnik, J. Application of PCA and time series analysis in studies of precipitation in Tricity (Poland). *Adv. Environ. Res.* **2004**, *8*, 337–349. [CrossRef]
4. Sayemuzzaman, M.; Jha, M.K. Seasonal and annual precipitation time series trend analysis in North Carolina, United States. *Atmos. Res.* **2014**, *137*, 183–194. [CrossRef]
5. Estévez, J.; Gavilán, P.; García-Marín, A.P.; Zardi, D. Detection of spurious precipitation signals from automatic weather stations in irrigated areas. *Int. J. Climatol.* **2015**, *35*, 1556–1568. [CrossRef]
6. Jiang, L.; Wu, J. *Hybrid PSO and GA for Neural Network Evolutionary in Monthly Rainfall Forecasting*; Springer: Berlin/Heidelberg, Germany, 2013; Volume 7802.

7. Cramer, S.; Kampouridis, M.; Freitas, A.A.; Alexandridis, A.K. An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives. *Expert Syst. Appl.* **2017**, *85*, 169–181. [CrossRef]
8. Teegavarapu, R.S.V.; Chandramouli, V. Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. *J. Hydrol.* **2005**, *312*, 191–206. [CrossRef]
9. Barrios, A.; Trincado, G.; Garreaud, R. Alternative approaches for estimating missing climate data: Application to monthly precipitation records in south-central Chile. *For. Ecosyst.* **2018**, *5*, 1–10. [CrossRef]
10. McCuen, R.H. *Hydrologic Analysis and Design*, 3rd ed.; Pearson: New York, NY, USA, 2004; ISBN 978-0131424241.
11. Bostan, P.A.; Heuvelink, G.B.M.; Akyurek, S.Z. Comparison of regression and kriging techniques for mapping the average annual precipitation of Turkey. *Int. J. Appl. Earth Obs. Geoinf.* **2012**, *19*, 115–126. [CrossRef]
12. Adhikary, S.K.; Muttil, N.; Yilmaz, A.G. Genetic Programming-Based Ordinary Kriging for Spatial Interpolation of Rainfall. *J. Hydrol. Eng.* **2016**, *21*, 04015062. [CrossRef]
13. Mair, A.; Fares, A. Comparison of Rainfall Interpolation Methods in a Mountainous Region of a Tropical Island. *J. Hydrol. Eng.* **2011**, *16*, 371–383. [CrossRef]
14. Simolo, C.; Brunetti, M.; Maugeri, M.; Nanni, T. Improving estimation of missing values in daily precipitation series by a probability density function-preserving approach. *Int. J. Climatol.* **2010**, *30*, 1564–1576. [CrossRef]
15. Xia, Y.; Fabian, P.; Stohl, A.; Winterhalter, M. Forest climatology: Estimation of missing values for Bavaria, Germany. *Agric. For. Meteorol.* **1999**, *96*, 131–144. [CrossRef]
16. Teegavarapu, R.S.V.; Tufail, M.; Ormsbee, L. Optimal functional forms for estimation of missing precipitation data. *J. Hydrol.* **2009**, *374*, 106–115. [CrossRef]
17. Teegavarapu, R.S.V. Estimation des données manquantes des précipitations en utilisant la proximité optimale d'imputation métrique base, la classification du plus proche voisin et méthodes d'interpolation à base de cluster. *Hydrol. Sci. J.* **2014**, *59*, 2009–2026. [CrossRef]
18. Huang, M.; Lin, R.; Huang, S.; Xing, T. A novel approach for precipitation forecast via improved K-nearest neighbor algorithm. *Adv. Eng. Inform.* **2017**, *33*, 89–95. [CrossRef]
19. Gorshenin, A.; Lebedeva, M.; Lukina, S.; Yakovleva, A. Application of Machine Learning Algorithms to Handle Missing Values in Precipitation Data. In *Lecture Notes in Computer Science*; (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Berlin, Germany, 2019; Volume 11965, pp. 563–577.
20. Bagirov, A.M.; Mahmood, A.; Barton, A. Prediction of monthly rainfall in Victoria, Australia: Clusterwise linear regression approach. *Atmos. Res.* **2017**, *188*, 20–29. [CrossRef]
21. Kajewska-Szkudlarek, J. Clustering approach to urban rainfall time series prediction with support vector regression model. *Urban Water J.* **2020**, *17*, 235–246. [CrossRef]
22. Estévez, J.; Bellido-Jiménez, J.A.; Liu, X.; García-Marín, A.P. Monthly Precipitation Forecasts Using Wavelet Neural Networks Models in a Semiarid Environment. *Water* **2020**, *12*, 1909. [CrossRef]
23. Partal, T.; Kişi, Ö. Wavelet and neuro-fuzzy conjunction model for precipitation forecasting. *J. Hydrol.* **2007**, *342*, 199–212. [CrossRef]
24. Li, G.; Ma, X.; Yang, H. A hybrid model for monthly precipitation time series forecasting based on variational mode decomposition with extreme learning machine. *Information* **2018**, *9*, 177. [CrossRef]
25. Filho, A.S.F.; Lima, G.A.R. Gap Filling of Precipitation Data by SSA—Singular Spectrum Analysis. *J. Phys. Conf. Ser.* **2016**, *759*, 012085.
26. Sun, M.; Li, X.; Kim, G. Precipitation analysis and forecasting using singular spectrum analysis with artificial neural networks. *Clust. Comput.* **2019**, *22*, 12633–12640. [CrossRef]
27. Kim, S.; Hong, S.; Joh, M.; Song, S.K. DeepRain: ConvLSTM network for precipitation prediction using multichannel radar data. *arXiv* **2017**, arXiv:1711.02316.
28. Ha, J.-H.; Lee, Y.H.; Kim, Y.-H. Forecasting the Precipitation of the Next Day Using Deep Learning. *J. Korean Inst. Intell. Syst.* **2016**, *26*, 93–98. [CrossRef]
29. Chen, L.; Cao, Y.; Ma, L.; Zhang, J. A Deep Learning-Based Methodology for Precipitation Nowcasting with Radar. *Earth Space Sci.* **2020**, *7*, e2019EA000812. [CrossRef]
30. Estévez, J.; Gavilán, P.; García-Marín, A.P. Spatial regression test for ensuring temperature data quality in southern Spain. *Theor. Appl. Climatol.* **2018**, *131*, 309–318. [CrossRef]
31. Estévez Gualda, J.; Gavilán, P.; Giráldez, J.V. Guidelines on validation procedures for meteorological data from automatic weather stations. *J. Hydrol.* **2011**, *402*, 144–154. [CrossRef]
32. Shanker, M.S.; Hu, M.Y.; Hung, M.S. Effect of data standardization on neural network training. *Omega* **1996**, *24*, 385–397. [CrossRef]
33. Luna, A.M.; Lineros, M.L.; Gualda, J.E.; Giráldez Cervera, J.V.; Madueño Luna, J.M. Assessing the Best Gap-Filling Technique for River Stage Data Suitable for Low Capacity Processors and Real-Time Application Using IoT. *Sensors* **2020**, *20*, 6354. [CrossRef]
34. Bellido-Jiménez, J.A.; Estévez, J.; García-Marín, A.P. New machine learning approaches to improve reference evapotranspiration estimates using intra-daily temperature-based variables in a semi-arid region of Spain. *Agric. Water Manag.* **2020**, *245*, 106558. [CrossRef]

35. Banadkooki, F.B.; Ehteram, M.; Ahmed, A.N.; Fai, C.M.; Afan, H.A.; Ridwam, W.M.; Sefelnasr, A.; El-Shafie, A. Precipitation forecasting using multilayer neural Network and support vector machine optimization based on flow regime algorithm taking into Account uncertainties of soft computing models. *Sustainability* **2019**, *11*, 6681. [CrossRef]
36. Ortiz-García, E.G.; Salcedo-Sanz, S.; Casanova-Mateo, C. Accurate precipitation prediction with support vector classifiers: A study including novel predictive variables and observational data. *Atmos. Res.* **2014**, *139*, 128–136. [CrossRef]
37. Nayak, M.A.; Ghosh, S. Prediction of extreme rainfall event using weather pattern recognition and support vector machine classifier. *Theor. Appl. Climatol.* **2013**, *114*, 583–603. [CrossRef]
38. Aftab, S.; Ahmad, M.; Hameed, N.; Bashir, M.S.; Ali, I.; Nawaz, Z. Rainfall prediction in Lahore City using data mining techniques. *Int. J. Adv. Comput. Sci. Appl.* **2018**, *9*, 254–260. [CrossRef]
39. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
40. Sukovich, E.M.; Ralph, F.M.; Barthold, F.E.; Reynolds, D.W.; Novak, D.R. Extreme quantitative precipitation forecast performance at the weather prediction center from 2001 to 2011. *Weather Forecast.* **2014**, *29*, 894–911. [CrossRef]
41. Das, S.; Chakraborty, R.; Maitra, A. A random forest algorithm for nowcasting of intense precipitation events. *Adv. Space Res.* **2017**, *60*, 1271–1282. [CrossRef]
42. Wolfensberger, D.; Gabella, M.; Boscacci, M.; Germann, U.; Berne, A. RainForest: A random forest algorithm for quantitative precipitation estimation over Switzerland. *Atmos. Meas. Tech.* **2021**, *14*, 3169–3193. [CrossRef]
43. Wang, Y.; Liu, J.; Li, R.; Suo, X.; Lu, E. Precipitation forecast of the Wujiang River Basin based on artificial bee colony algorithm and backpropagation neural network. *Alex. Eng. J.* **2020**, *59*, 1473–1483. [CrossRef]
44. Kotthoff, L.; Thornton, C.; Hoos, H.; Hutter, F.; Leyton-Brown, K. Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. *J. Mach. Learn. Res.* **2017**, *18*, 1–5.
45. Jin, H.; Song, Q.; Hu, X. Auto-Keras: An Efficient Neural Architecture Search System. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 1946–1956.
46. Feurer, M.; Klein, A.; Eggensperger, K.; Springenberg, J.T.; Blum, M.; Hutter, F. Auto-sklearn: Efficient and robust automated machine learning. In Proceedings of the 28th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; Volume 2, pp. 2962–2970.
47. Hutter, F.; Kotthoff, L.; Vanschoren, J. *Automated Machine Learning*; The Springer Series on Challenges in Machine Learning; Springer International Publishing: Cham, Switzerland, 2019; ISBN 978-3-030-05317-8.
48. Bellido-Jiménez, J.A.; Estévez, J.; García-Marín, A.P. Assessing Neural Network Approaches for Solar Radiation Estimates Using Limited Climatic Data in the Mediterranean Sea. In Proceedings of the 3rd International Electronic Conference on Atmospheric Sciences (ECAS 2020), Online, 16–30 November 2020.
49. Borji, A.; Itti, L. Bayesian optimization explains human active search. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 55–63.
50. Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R.P.; de Freitas, N. Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proc. IEEE* **2016**, *104*, 148–175. [CrossRef]
51. Kim, J.; Ryu, J.H. A heuristic gap filling method for daily precipitation series. *Water Resour. Manag.* **2016**, *30*, 2275–2294. [CrossRef]
52. Wuthiwongyothin, S.; Kalkan, C.; Panyavaraporn, J. Evaluating Inverse Distance Weighting and Correlation Coefficient Weighting Infilling Methods on Daily Rainfall Time Series. *SNRU J. Sci. Technol.* **2021**, *13*, 71–79.
53. Sehad, M.; Lazri, M.; Ameur, S. Novel SVM-based technique to improve rainfall estimation over the Mediterranean region (North of Algeria) using the multispectral MSG SEVIRI imagery. *Adv. Space Res.* **2017**, *59*, 1381–1394. [CrossRef]