water

MDPI

Article

Artificial Neural Networks and Multiple Linear Regression for Filling in Missing Daily Rainfall Data

Ioannis Papailiou [1], Fotios Spyropoulos [1], Ioannis Trichakis [2,*] and George P. Karatzas [1]

[1] School of Chemical and Environmental Engineering, Technical University of Crete, 73100 Chania, Greece
[2] European Commission, Joint Research Centre, 21027 Ispra, Italy
* Correspondence: ioannis.trichakis@ec.europa.eu

Abstract: As demand for more hydrological data has been increasing, there is a need for the development of more accurate and descriptive models. A pending issue regarding the input data of said models is the missing data from observation stations in the field. In this paper, a methodology utilizing ensembles of artificial neural networks is developed with the goal of estimating missing precipitation data in the extended region of Chania, Greece on a daily timestep. In the investigated stations, there have been multiple missing data events, as well as missing data prior to their installation. The methodology presented aims to generate precipitation time series based on observed data from neighboring stations and its results have been compared with a Multiple Linear Regression model as the basis for improvements to standard practice. For each combination of stations missing daily data, an ensemble has been developed. According to the statistical indexes that were calculated, ANN ensembles resulted in increased accuracy compared to the Multiple Linear Regression model. Despite this, the training time of the ensembles was quite long compared to that of the Multiple Linear Regression model, which suggests that increased accuracy comes at the cost of calculation time and processing power. In conclusion, when dealing with missing data in precipitation time series, ANNs yield more accurate results compared to MLR methods but require more time for producing them. The urgency of the required data in essence dictates which method should be used.

Keywords: rainfall time series; artificial neural networks; Multiple Linear Regression; Chania

Citation: Papailiou, I.; Spyropoulos, F.; Trichakis, I.; Karatzas, G.P. Artificial Neural Networks and Multiple Linear Regression for Filling in Missing Daily Rainfall Data. *Water* **2022**, *14*, 2892. https://doi.org/10.3390/w14182892

Academic Editors: Fi-John Chang, Li-Chiu Chang and Jui-Fa Chen

Received: 26 August 2022
Accepted: 13 September 2022
Published: 16 September 2022

1. Introduction

The successful development of reliable models for predicting the status of water resources of a particular region is inextricably linked to the quantity and quality of the climate and hydrological data used [1]. One of the most critical pieces of data for such a study is the available rainfall data in the area of interest [2]. The possibility of errors or gaps within an available rainfall data time series is real and may be due to errors in the measuring instruments, a possible instrument failure, or an extreme weather event. Therefore, the development of a model capable of accurately simulating, or even complementing, a time series of rainfall data is necessary.

The importance of rainfall data availability is inarguable in hydrological modelling as these data are an essential input parameter in almost any approach. Previous research has supported the notion that the traditional statistical methods for infilling (imputing) missing data may be inefficient for small temporal and spatial scales [3,4]. Thus, an indicator of the success of the model is its outperformance over standard interpolation methods. Such practices have become more nuanced over the years, specifically with the incorporation of weighting factors that compensate for the variation between stations due to the morphological features of each case study [5].

When looking at the recently published scientific literature, Artificial Neural Networks have shown encouraging results in modeling nonlinear problems, such as hydrological processes [6]. They are able to recognize strong seasonal patterns without the need for

preprocessing raw data to remove outliers, and there is solid evidence that supports the accuracy of their prediction [7]. A work similar to the current article has been conducted using meteorological data from the internet, with the intent of forecasting future rainfall using multi-layer perceptron (MLP) with back propagation and optimization algorithms [8]. In another work, the MLPs are used for forecasting future precipitation using rainfall data from nearby weather stations as inputs [9]. As an alternative method for monthly rainfall prediction, it has been suggested that the use of ANNs with wavelet regression provides more accurate results compared to models using ANNs, which implies the need for optimization [10]. An alternative to MLPs is Long Short-Term Memory networks, which are a class of recurrent neural networks that have shown promising results in estimating runoff from rainfall. With respect to the problem at hand, the selected neural networks provide a high degree of regression ability. Using recurrent networks, like those used in rainfall runoff modelling [11], would not have a physical meaning, since the relationship between inputs and outputs (daily rainfall values) does not include a temporal delay. Other techniques for filling in missing data in the field of hydrology include K-nearest neighbors (KNN), adaptive neuro-fuzzy inference systems (ANFISs) and random forest regression (RFR) [12–14], but these go beyond the scope of this work and could be considered for future research. Regarding the number of inputs, large numbers of different inputs do not guarantee more accurate results. A genetic algorithm can improve the process of selection when aiming for forecasting, but in this work, in order to reduce computational demands and given the nature of the network, another optimization method was chosen [15]. Apart from genetic algorithms as optimization techniques, others exist, such as particle swarm, cuckoo search, and bat- or kidney-inspired algorithms, depending on the level of strictness demanded [16]. In this paper, optimization is achieved through the use of a competitive algorithm in the creation of each ensemble, corresponding to each combination of missing data from the observation stations. Artificial intelligence tools have been implemented in the past in different scientific fields, from filling in spatially and temporally missing data by using augmented interpolation [17] to using photonic neural networks analysis for the changing morphology of an area [18]. In regions with high unpredictability due to extreme weather conditions, ANNs have been successful in forecasting rainfall [19]; given this fact, ANNs might perform even better in regions with strong seasonal patterns and a temperate climate, such as Crete. In large areas with varied topography, proximity of stations does not always guarantee a correlation between observed rainfall values, especially if the stations belong in two different hydrological catchments [1]. In the current case study, the area is hydrologically homogenous with only a small increase in precipitation at higher elevations [20]. In addition, fluctuations between extreme values can be smoothed out by classifying data either spatially [21] or based on intensity [22], which implies training and using multiple ANNs. Multiple ANNs with targeted training working on their own niche outperform an all-purpose ANN trained with the whole data set, with differences being dependent on the physical problem [23]. As hinted previously, multiple ANNs creating an ensemble might outperform a singular one by minimizing the occurrence of local minimums and individual biases [24]. The most simplified approach to composing an ensemble of neural networks is averaging their results using simple or weighted averages. Previous research has also proposed that the structure of the ANN ensemble can itself become the input of a general regression neural network [25]. This technique can exploit the variability of results produced by biased individuals and increase overall accuracy. In addition, it utilizes a full set of ANNs in which there may be individuals that produce error-increasing results. In order to address this, it is suggested to develop competitive algorithms where ANNs or ensembles are compared to each other and the best-performing method ends up being used for predictions [26]. In the same manner of thinking, elimination of the least significant input variables can be performed in an ensemble by considering the correlation coefficient, which has been mostly applied to climatic variables in forecasting rather than regression-based forecasting [27]. One approach to creating an ensemble with a limited data set is to alternate between training and testing data sets during the training

period and eventually average out the ensemble outputs [28]. Another issue arising when working with ANNs, especially ensembles, is the network architecture, since it can greatly impact the performance; in most cases, an optimization algorithm is developed since there is no standard and optimal architecture is defined by trial and error [24]. Finally, one optimization technique which borders on architecture modification is the dropout method which randomly turns off units and their connections during training [29], which shows that random-based optimization might produce adequate results.

This paper aims to develop a methodology to estimate missing daily precipitation values from weather stations. Five weather stations monitoring rainfall in the prefecture of Chania, Greece, were used as a case study. This work focuses on the comparison of ANN ensembles based on multi-layer perceptrons and the more commonly used multiple linear regression (MLR) for completion of time series of daily rainfall data. This way, the results of the ANNs are compared to a technique that is standard practice in the field (MLR) [13]. In this approach, the best ANN from each ensemble imputes the missing data values to end up with a completed dataset for all stations. It is important to state that classification based on different combinations of missing data (henceforth called cases) adds to the accuracy of the model in general, since the ANNs are specialized in each case. This would not be feasible if modeling was done by creating a single ensemble for all stations, or an ensemble for each station. The respective MLR results are calculated as a baseline for comparison.

2. Materials and Methods

2.1. ANN and MLR Creation

The proposed methodology starts from a dataset with missing rainfall data for some stations and results in two completed datasets from the ANN ensembles and the MLR. The first step of the algorithm is to check every date containing recorded data. If a daily dataset has no missing values, then it is included in the dataset which will be used for training and validation of the ANN ensembles and validation of the MLR model. Otherwise, it is added to the dataset meant for imputing. It is important at this point to state that if a daily dataset has no recorded data at any of the stations, then imputation is unfeasible with the proposed methodology, primarily because completion of the time series occurs on a daily timestep by correlating the missing data with the observed data. In addition, a precipitation event is not dependent on a past precipitation event, and since rainfall is the sole input in this model, it was deemed both unnecessary and accuracy-decreasing to impute the time series by correlating data from datasets that correspond to different dates. This is the reason why the completed time series span from the first recorded dataset up to the current day and not further into the past or future.

The outcomes of the separation are two datasets: a complete and an incomplete one. The full daily datasets will be used for the training and validation of the ANN ensemble. Due to the different cases of missing data, it was deemed necessary to create multiple ANNs (multiple layer perceptron) that are specialized to each case, since inputs and outputs for each case differ, which implies a different topology for each case. The inputs and outputs are always daily rainfall values from weather stations, and for each different combination of missing data, the stations with observed values are used as input nodes and the stations with missing values are used as output nodes. In order to increase the accuracy of the model altogether, for each case an ensemble of 10,000 ANNs with one hidden layer was trained, in which the daily datasets for training and validation were randomly selected from the full set. With the use of a competitive algorithm, only one ANN—the best-performing one, according to its test error value—was selected to give outputs, using MATLAB's ANN tool version 2017b. According to the literature [30,31], one hidden layer is sufficient and might also outperform ANNs with multiple hidden layers when used for regression. The competitive algorithm selects the best-performing ANN based on training error and the results are produced solely based on that ANN. The use of ensembles instead of one single ANN addresses any concerns regarding the reliability, performance, and behavior of the proposed approach. The calibration (training and validation) dataset was 80% of the full

available dataset with complete records, and the testing dataset consisted of the remaining 20% for all ANNs. After the training and validation are conducted, the ensembles are ready to complete the time series. Similarly, the MLR functions are created by the training and validation dataset for each case. After both processes have completed the time series, all negative values that are generated are turned into zeroes.

The whole process is graphically represented in Figure 1 below.

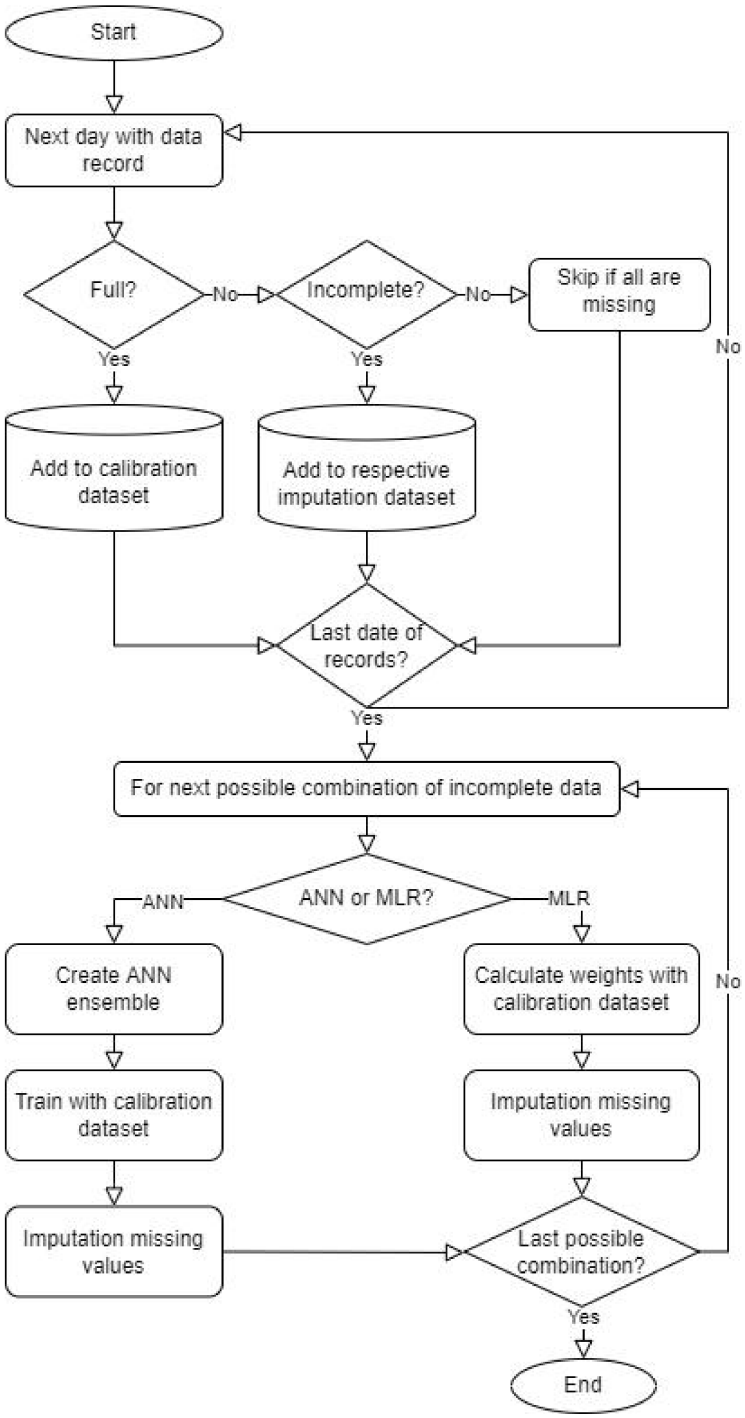

Figure 1. Flowchart of the methodology.

2.2. Model Evaluation

The validity of the results of both models is verified by the calculation of the correlation coefficients between the target and the simulated value. The value of the Nash–Sutcliffe

coefficient is calculated, which can take values from minus infinity to one ($-\infty$ to 1), based on which the validity of the model is determined, with a value of one (1) indicating complete agreement between the simulated values given by the model and those observed by the stations. According to the literature, an NSE index value above 0.7 corresponds to a very good estimation [32]. Finally, the Root Mean Square Error is extracted from the model results in each of the cases considered [32].

2.3. Case Study

In the prefecture of Chania, near the northern coast of Crete, there are five automatic weather stations at a relatively close distance (approximately 5 km) to each other, as shown in Figure 2.

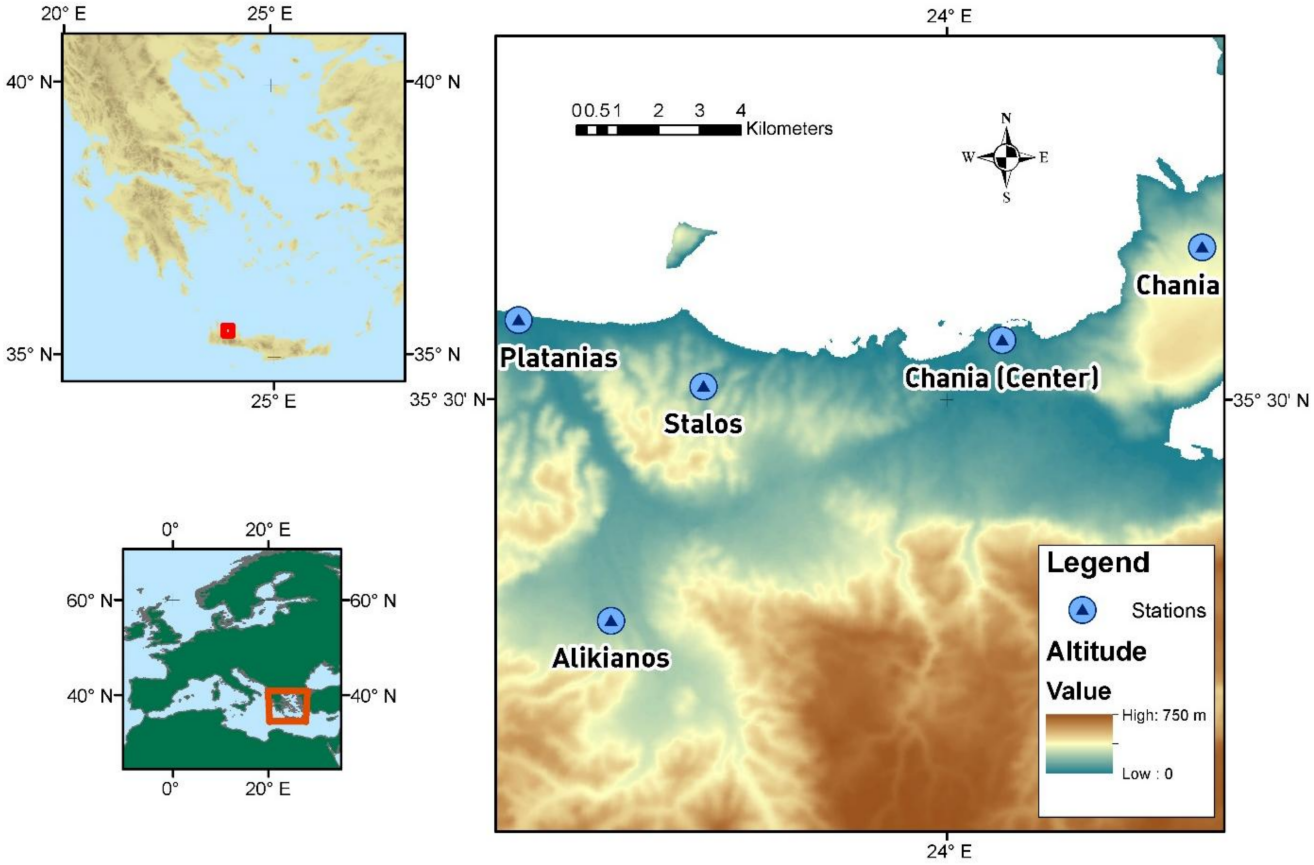

Figure 2. Weather station locations.

Regarding the locations of the stations as shown in Figure 2, the overall highest value of rainfall, historically, has occurred at Alikianos station, while the lowest has occurred at Platanias station. Platanias station has the lowest recorded altitude at 12 m, while Alikianos station is located at 95 m. Chania station (137 m) is located at a higher altitude than Alikianos station (95 m). Although it would be expected that a station at a higher altitude has a greater amount of rainfall, it was observed from the data that Alikianos station has a greater amount of rainfall. The reason for this might be that Alikianos station is furthest from the sea compared to all the other stations considered and is situated at the foot of the Lefka Ori. Platanias, on the other hand, is located a short distance from the sea and at a low altitude.

Table 1 contains a summary of the recorded daily precipitation values available from the automatic weather station NOANN network [33] (in total 15,040 records).

Table 1. Daily data availability and initial operating day of each rainfall station.

Station	Altitude (m)	Number of Data	Start of Data
Alikianos	95	3044	1 September 2012
Chania	137	5448	1 February 2006
Chania (Center)	7	3745	1 October 2010
Platanias	12	2011	1 July 2015
Stalos	93	792	1 November 2018

Based on these records, a timeline showing the availability and gaps in the datasets for the study period is shown in Figure 3. In total, 759 days had a complete dataset and were used for calibration and 4689 days had at least one missing value.

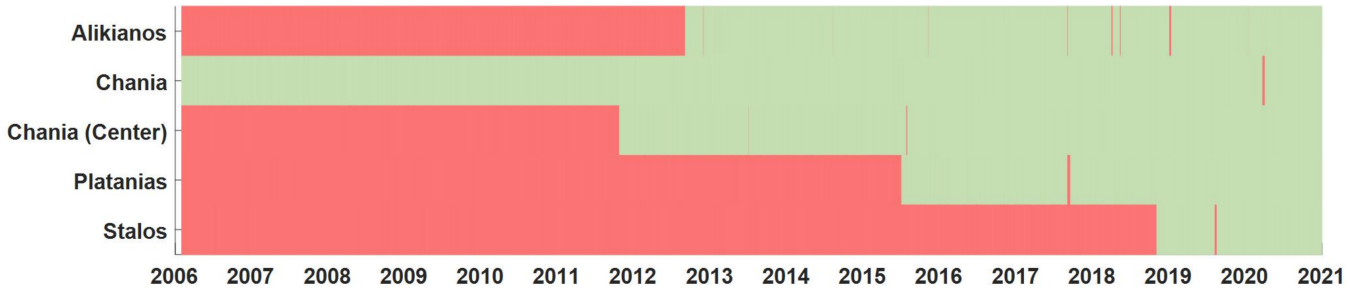

Figure 3. Timeline of daily rainfall data availability and gaps in the datasets (red color indicates gaps).

The recording of the data used in this work starts with the creation of Chania station on 1 February 2006. This means that for the period from 1 February 2006 to 30 September 2010, the available rainfall data originates only from Chania station. As of the next day, on 1 October 2010, when Chania station (Center) was put into operation, the recorded rainfall data come from the two stations previously mentioned. On 1 September 2012 the Alikianos meteorological station was put into function, therefore the recorded rainfall data come from the above three meteorological stations. To continue, on 1 July 2015, the recording of rainfall data from Platanias meteorological station starts, which means that the model input data comes from four stations. Finally, on 1 November 2018, the last station, Stalos, was put into operation. Therefore, for the next period, we have logging data from all five stations until 31 December 2020. It is worth noting that the period of time that a station is in operation is not always the same as the period of time that it records data, as there may be losses due to errors in the measuring instruments, a possible instrument failure, or an extreme weather phenomenon. This is clearly shown in Figure 3 of the paper.

2.4. Different Combinations of Stations Missing Data (Cases)

There are five rainfall stations in our study and each one of them has a different installation date, from which point on data are available. In addition, there are periods when, for different reasons (maintenance, power cuts, malfunction), one or more daily values are missing from the time series. The values missing for each day, together with the values available, can be categorized into different cases, in order to organize and group the different dates based on different calculation needs.

Figure 4 shows all the possible combinations of stations having or missing a daily record. By having all the possible cases identified, the algorithm is able to create ensembles for cases that have not occurred yet.

In the full, observed dataset, 9 cases occur out of a total of 32 that were theoretically possible. Specifically, the included cases are Case 2, Case 3, Case 6, Case 11, Case 14, Case 15, Case 22, Case 24, and Case 29. In three cases, namely Cases 2, 3 and 6, one station had a missing value; in three other cases, namely Cases 11, 14 and 15, two stations had missing values; in two cases, Cases 22 and 24, three stations had missing values; and in the last case, Case 29, four stations had missing values. The numbering of each case is not derived from

the numerical order, but from the corresponding case, as shown in Figure 4. For example, in Case 2 the input precipitation data are the values from Chania, Chania (Center), Platanias and Stalos stations, and the output is the precipitation value for Alikianos station.

Case	Alikianos	Chania	Chania (Center)	Platanias	Stalos
1	1	1	1	1	1
2	**0**	**1**	**1**	**1**	**1**
3	**1**	**0**	**1**	**1**	**1**
4	1	1	0	1	1
5	1	1	1	0	1
6	**1**	**1**	**1**	**1**	**0**
7	0	0	1	1	1
8	0	1	0	1	1
9	1	0	1	1	0
10	0	1	1	0	1
11	**0**	**1**	**1**	**1**	**0**
12	1	0	0	1	1
13	1	1	0	0	1
14	**1**	**1**	**1**	**0**	**0**
15	**1**	**1**	**0**	**1**	**0**
16	1	0	1	0	1
17	0	0	0	1	1
18	0	0	1	0	1
19	0	0	1	1	0
20	0	1	0	0	1
21	0	1	0	1	0
22	**0**	**1**	**1**	**0**	**0**
23	1	0	0	0	1
24	**1**	**1**	**0**	**0**	**0**
25	1	0	1	0	0
26	1	0	0	1	0
27	0	0	0	0	1
28	1	0	0	0	0
29	**0**	**1**	**0**	**0**	**0**
30	0	0	1	0	0
31	0	0	0	1	0
32	0	0	0	0	0

Figure 4. Possible combinations of availability of daily rainfall data. Red indicates that the station in question has no recorded rainfall value for the day of recording. Cases occurring in the dataset are shown in bold.

3. Results

After completing a full run of the algorithm built using the proposed methodology, the incomplete time series of each station receives model-generated data for the full period in which at least one of the five stations has an observed value. In the following charts (Figure 5), the results of the two methodologies are shown for all stations. In the left column, the model-generated values of the ANN have an orange color, and in the right column, the model-generated values of the MLR have a red color, while the observed values in all charts are in a blue color.

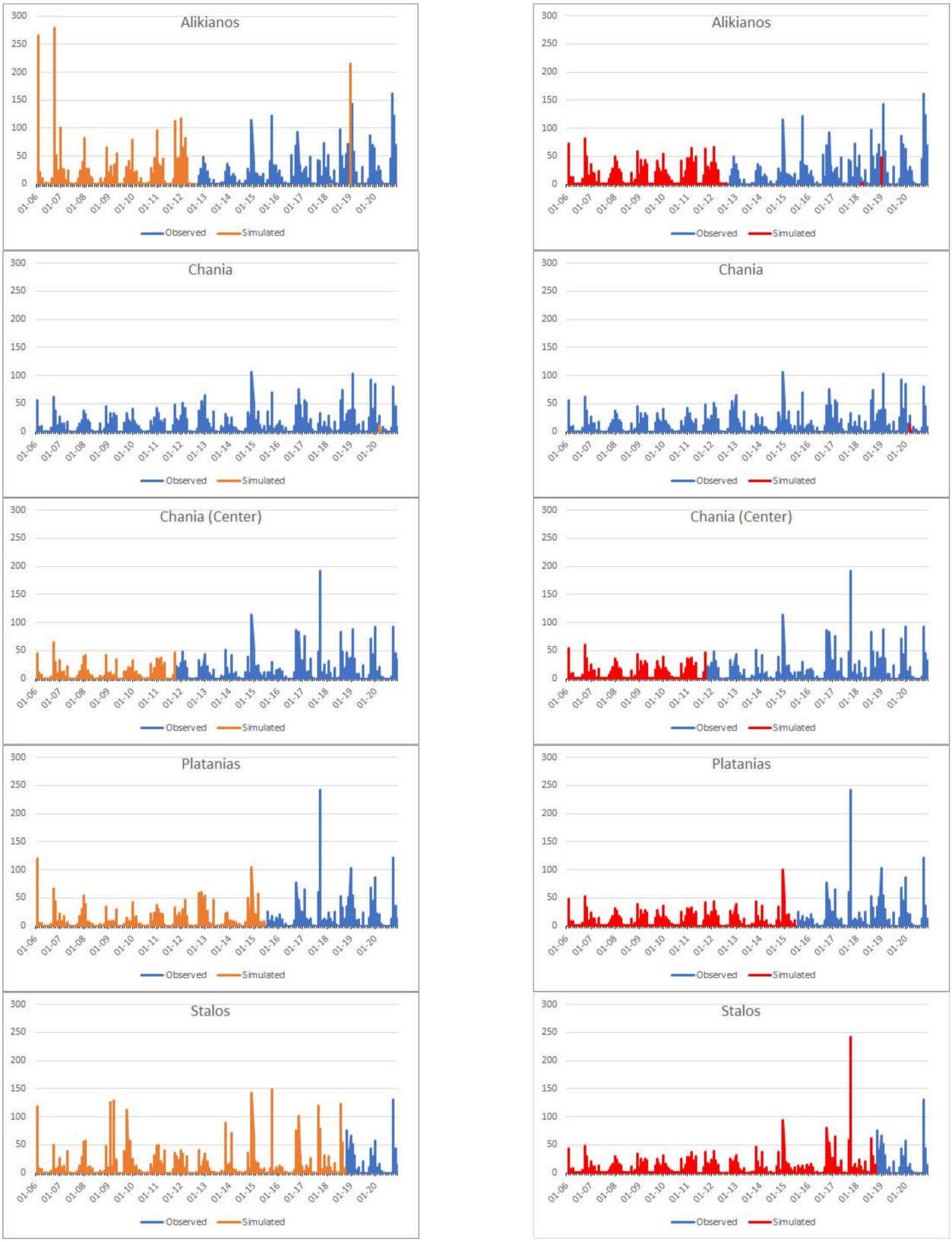

Figure 5. Observed values (blue), and model-generated values from the ANNs (orange) and MLR model (red).

To compare the two methods, three different metrics were used, the root mean square error, the Nash–Sutcliffe efficiency coefficient, and the correlation coefficient. The results

are shown on a per-case basis, as the two methods might show different sensitivities to missing data. Comparative tables at the end of each section summarize the results of the testing dataset.

Concerning the computational effort and time needed for the two methods, the ANN did take a considerably large amount of time to optimize its structure (almost 36 h on a PC (Personal Computer) with an Intel i7 8th generation processor). On the other hand, the MLR was significantly faster, and required only a few minutes to run.

3.1. Root Mean Square Error (RMSE)

The RMSE index indicates the deviation between the observed and simulated values and indicates whether the data are clustered around the line of best fit. The models calculate the Root Mean Square Error (RMSE) for each of the cases considered. Regarding the Artificial Neural Network model, the best value is presented in Case 15 and shows an error equal to 1.16 mm, while the worst value is presented in Case 29, with an error value equal to 2.42 mm. The corresponding results of the Multiple Linear Regression model are shown in Case 3 with a value of 2.37 mm and Case 2 with a value of 6.43 mm. Overall, the Artificial Neural Network model shows lower errors, ranging from 42% to 72.6%, compared to the Multiple Linear Regression model.

The following Table 2 contains all the above results aggregated as follows:

Table 2. Root Mean Square Error of testing dataset.

Case	RMSE [mm]	
	ANN	MLR
Case 2	1.76	6.43
Case 3	1.22	2.37
Case 6	1.22	2.92
Case 11	2.30	4.99
Case 14	1.24	3.03
Case 15	1.16	2.46
Case 22	2.19	4.47
Case 24	1.83	3.16
Case 29	2.42	4.94

3.2. Nash–Sutcliffe Efficiency

The Nash–Sutcliffe coefficient can take values from minus infinity to one (-∞ to 1), where for these values the following applies:

- If NSE = 1, then there is a complete match between the simulated values given by the model and those observed by the stations;
- If NSE = 0, then the values simulated by the model give the same result as if the average of the observed values of the stations were used as the forecast model for each time point;
- If NSE < 0, then the model is practically unusable, as the values simulated by it give a less accurate result than if the average of the observed values of the stations were used as a predictive model for each time point.

With respect to the calculation of the Nash–Sutcliffe coefficients, the Artificial Neural Network model shows, again, higher overall values ranging from 2.1% to 28.7%. For the Artificial Neural Network model, the best value of the Nash–Sutcliffe coefficient is presented in Case 15, with a value of 0.989, while the worst value is presented in Case 29, with a value of 0.911. The corresponding results for the Multiple Linear Regression model appear in Case 15 with a value of 0.968 and in Case 29 with a value of 0.708.

Similarly, the Nash–Sutcliffe Efficiency values for all cases are presented in the following Table 3 which contains all the results in an aggregated way:

Table 3. Nash–Sutcliffe Efficiencies of testing dataset.

Case	Nash–Sutcliffe Efficiency		Simulated Precipitation Value Station(s)
	ANN	MLR	
Case 2	0.967	0.803	Alikianos
Case 3	0.975	0.937	Chania
Case 6	0.981	0.882	Stalos
Case 11	0.954	0.803	Alikianos
	0.957	0.882	Stalos
Case 14	0.976	0.908	Platanias
	0.969	0.845	Stalos
Case 15	0.989	0.968	Chania (Center)
	0.973	0.871	Stalos
Case 22	0.934	0.802	Alikianos
	0.957	0.908	Platanias
	0.927	0.844	Stalos
Case 24	0.975	0.954	Chania (Center)
	0.957	0.869	Platanias
	0.943	0.781	Stalos
Case 29	0.911	0.708	Alikianos
	0.971	0.933	Chania (Center)
	0.968	0.843	Platanias
	0.959	0.748	Stalos

The results show a clear increase in the performance of the Nash–Sutcliffe efficiency when using the ANN instead of MLR. The ANN's performance was also higher when fewer stations were available compared to its MLR counterpart, which had a declining performance especially when one or two stations were available. It is also clear that there is a great correlation between the Chania (Center) and Chania stations, so when one is available, the results for the other are always very good. This is confirmed by the results of Case 3 where only Chania station is missing and from the results of Cases 15, 24 and 29, where station Chania is available, and Chania (Center) is missing.

3.3. Coefficient of Correlation (R)

The Coefficient of Correlation (R) indicates the proportion of variance of the dependent variable derived from the independent variable. A value of one (1) is the maximum value the coefficient can take, which indicates that there is a complete match between the two compared values.

Regarding the calculation of the Correlation Coefficient (R) for each case, the Artificial Neural Network model shows higher overall values ranging from 5.4% to 29.7%. More specifically, the best value of the above coefficient for the Artificial Neural Network model is presented in Case 15, with a value of 0.99274, while the worst value is presented in Case 6, with a value of 0.93957. The corresponding results for the Multiple Linear Regression model appear in Case 3, with a value of 0.93740 and in Case 29, with a value of 0.74782. Similarly, the Coefficients of Correlation for each case are presented in the following Table 4 which contains the aggregated results:

Table 4. Coefficients of Correlation of testing dataset.

Case	Coefficient of Correlation (R)	
	ANN	MLR
Case 2	0.98353	0.80337
Case 3	0.98777	0.93740
Case 6	0.99066	0.88198
Case 11	0.97737	0.76844
Case 14	0.98639	0.87493
Case 15	0.99101	0.90800
Case 22	0.96842	0.78287
Case 24	0.97975	0.86749
Case 29	0.96998	0.74782

4. Discussions

This work developed and compared two models for the simulation of precipitation values, which simulated and accurately completed five time series of precipitation data from five meteorological stations in the region of Chania, Crete. The first model was developed using an Artificial Neural Network ensemble approach (similar to other previously published works [6,10,27]), while the second model was developed using the Multiple Linear Regression method, both in a MATLAB environment.

It is observed that the four meteorological stations that are relatively close to the sea, while at the same time are relatively close to each other (Chania, Chania (Centre), Platanias and Stalos), show similar results for their total rainfall values (Figure 5). From a hydrological standpoint, both models present results that are in accordance with the theoretical expectations; the simulated values at the weather stations near the seafront are always lower when compared to those of stations at higher altitude. In addition, there is a small decline in the precipitation values along the west to east axis, which is expected since most of the water load in the clouds is released when they reach the coastal fronts coming from the Western Mediterranean.

Looking at the ANN results, a couple of simulated values might draw the attention of the reader as being exceptionally high and possibly outliers (e.g., October 2006 and January 2019). Nevertheless, the scientific literature and the observed values from already installed stations confirm that these were months with extreme rainfall events, confirming the plausibility of these simulated values. In October 2006, extreme rainfall events occurred throughout the study area, leading to flooding in the city of Chania, serious material damages and one casualty [20]. At that time, the only installed and operating station was the one in Chania, which had a very high observed value of 214.6 mm, one of the highest ever recorded. For the same month, the simulated precipitation value for Alikianos station is 345 mm based on ANNs, while the corresponding value using the MLR method is 194 mm. These values, although they seem quite high for the area concerned, are in accordance with the value observed in Chania. In January and February 2019, other extreme rainfall events occurred with similar results. In 2019, all weather stations were operational, but there was a 10-day gap in the beginning of January in Alikianos station, possibly because of device failure due to the extremity of the rainfall events. Regarding the month of January 2019, the simulated precipitation value for Alikianos station is 692 mm based on ANNs, while the corresponding value using the MLR method is 362 mm. The extremity of those values is confirmed by the literature, while the events continued in February with the Chionis and Oceanida storms [20]. The seemingly high simulated value for January is confirmed by the observed values in February at all weather stations. In Figure 5, the observed values in February are significantly high, with the highest value recorded at Alikianos station (568.8 mm in total) and the next highest value at Chania station (360 mm in total). Based on the above, we conclude that the simulated values for Alikianos are plausible and do not consider them as outliers. Comparing the two models, the results of the ANN model show that it is more capable of simulating extreme weather values compared to the model obtained with the MLR method.

5. Conclusions

Both methods have proven more than adequate for the task of imputation of gaps in the daily rainfall time series. The Nash–Sutcliffe coefficient for both methods is above 0.7 for all cases, a value generally considered as the threshold for very good performance. Nevertheless, throughout this work, the Artificial Neural Network ensembles consistently outperformed the Multiple Linear Regression model. The obvious caveat is the increased time needed for training the ANN model. When comparatively small datasets are available for training (like in this work), the computational effort for training the ANN ensembles is also relatively small (taking just over thirty-six hours). In such cases using ANNs might make more sense, always considering the urgency of the application. In cases where the available dataset is large the training time is expected to increase, but the results will

probably be better than those obtained with Multiple Linear Regression. A decision should be made as to whether accuracy or speed is more important. For increased accuracy, the results of this study suggest using ANNs, for increased speed, the results point to using Multiple Linear Regression. Given the good performance of the ensembles in this work, future work can focus on testing different activation functions like the reLU and tanhLU [34].

Author Contributions: Conceptualization, I.T.; methodology, I.T.; software, I.P. and I.T.; validation, I.P. and I.T.; data curation, I.P.; writing—original draft preparation, I.P. and F.S.; writing—review and editing, F.S. and I.T.; visualization, I.T.; supervision, G.P.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

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

References

1. Canchala-Nastar, T.; Carvajal-Escobar, Y.; Alfonso-Morales, W.; Loaiza Cerón, W.; Caicedo, E. Estimation of Missing Data of Monthly Rainfall in Southwestern Colombia Using Artificial Neural Networks. *Data Brief* **2019**, *26*, 104517. [CrossRef] [PubMed]
2. Nkuna, T.R.; Odiyo, J.O. Filling of Missing Rainfall Data in Luvuvhu River Catchment Using Artificial Neural Networks. *Phys. Chem. Earth Parts A/B/C* **2011**, *36*, 830–835. [CrossRef]
3. Tran Anh, D.; Van, S.P.; Dang, T.D.; Hoang, L.P. Downscaling Rainfall Using Deep Learning Long Short-term Memory and Feedforward Neural Network. *Int. J. Climatol.* **2019**, *39*, 4170–4188. [CrossRef]
4. Ben Aissia, M.-A.; Chebana, F.; Ouarda, T.B. Multivariate Missing Data in Hydrology—Review and Applications. *Adv. Water Resour.* **2017**, *110*, 299–309. [CrossRef]
5. Teegavarapu, R.S.V.; Aly, A.; Pathak, C.S.; Ahlquist, J.; Fuelberg, H.; Hood, J. Infilling Missing Precipitation Records Using Variants of Spatial Interpolation and Data-Driven Methods: Use of Optimal Weighting Parameters and Nearest Neighbour-Based Corrections: INFILLING MISSING PRECIPITATION RECORDS. *Int. J. Climatol.* **2018**, *38*, 776–793. [CrossRef]
6. Elshaboury, N.; Elshourbagy, M.; Al-Sakkaf, A.; Abdelkader, E.M. Rainfall Forecasting in Arid Regions Using an Ensemble of Artificial Neural Networks. *J. Phys. Conf. Ser.* **2021**, *1900*, 012015. [CrossRef]
7. Mishra, N.; Soni, H.K.; Sharma, S.; Upadhyay, A.K. A Comprehensive Survey of Data Mining Techniques on Time Series Data for Rainfall Prediction. *J. ICT Res. Appl.* **2017**, *11*, 167–183. [CrossRef]
8. Kashiwao, T.; Nakayama, K.; Ando, S.; Ikeda, K.; Lee, M.; Bahadori, A. A Neural Network-Based Local Rainfall Prediction System Using Meteorological Data on the Internet: A Case Study Using Data from the Japan Meteorological Agency. *Appl. Soft Comput.* **2017**, *56*, 317–330. [CrossRef]
9. Ridwan, W.M.; Sapitang, M.; Aziz, A.; Kushiar, K.F.; Ahmed, A.N.; El-Shafie, A. Rainfall Forecasting Model Using Machine Learning Methods: Case Study Terengganu, Malaysia. *Ain Shams Eng. J.* **2021**, *12*, 1651–1663. [CrossRef]
10. Goyal, M.K. Monthly Rainfall Prediction Using Wavelet Regression and Neural Network: An Analysis of 1901–2002 Data, Assam, India. *Theor. Appl. Climatol.* **2014**, *118*, 25–34. [CrossRef]
11. Hu, C.; Wu, Q.; Li, H.; Jian, S.; Li, N.; Lou, Z. Deep Learning with a Long Short-Term Memory Networks Approach for Rainfall-Runoff Simulation. *Water* **2018**, *10*, 1543. [CrossRef]
12. Qin, Y.; Lou, Y. Hydrological Time Series Anomaly Pattern Detection Based on Isolation Forest. In Proceedings of the 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chengdu, China, 15–17 March 2019; pp. 1706–1710.
13. Khazaee Poul, A.; Shourian, M.; Ebrahimi, H. A Comparative Study of MLR, KNN, ANN and ANFIS Models with Wavelet Transform in Monthly Stream Flow Prediction. *Water Resour. Manag.* **2019**, *33*, 2907–2923. [CrossRef]
14. Desai, S.; Ouarda, T.B. Regional Hydrological Frequency Analysis at Ungauged Sites with Random Forest Regression. *J. Hydrol.* **2021**, *594*, 125861. [CrossRef]
15. Haidar, A.; Verma, B. A Novel Approach for Optimizing Climate Features and Network Parameters in Rainfall Forecasting. *Soft Comput.* **2018**, *22*, 8119–8130. [CrossRef]
16. Jaddi, N.S.; Abdullah, S. Optimization of Neural Network Using Kidney-Inspired Algorithm with Control of Filtration Rate and Chaotic Map for Real-World Rainfall Forecasting. *Eng. Appl. Artif. Intell.* **2018**, *67*, 246–259. [CrossRef]
17. Cheng, S.; Lu, F. A Two-Step Method for Missing Spatio-Temporal Data Reconstruction. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 187. [CrossRef]
18. Yen, M.-H.; Liu, D.-W.; Hsin, Y.-C.; Lin, C.-E.; Chen, C.-C. Application of the Deep Learning for the Prediction of Rainfall in Southern Taiwan. *Sci. Rep.* **2019**, *9*, 12774. [CrossRef]
19. Lee, J.; Kim, C.-G.; Lee, J.; Kim, N.; Kim, H. Application of Artificial Neural Networks to Rainfall Forecasting in the Geum River Basin, Korea. *Water* **2018**, *10*, 1448. [CrossRef]

20. Goumas, C.; Trichakis, I.; Vozinaki, A.-I.; Karatzas, G.P. Flood Risk Assessment and Flow Modeling of the Stalos Stream Area. *J. Hydroinform.* **2022**, *24*, 677–696. [CrossRef]
21. Praveen, B.; Talukdar, S.; Shahfahad, M.S.; Mondal, J.; Sharma, P.; Islam, A.R.M.D.T.; Rahman, A. Analyzing Trend and Forecasting of Rainfall Changes in India Using Non-Parametrical and Machine Learning Approaches. *Sci. Rep.* **2020**, *10*, 10342. [CrossRef]
22. Beritelli, F.; Capizzi, G.; Lo Sciuto, G.; Napoli, C.; Scaglione, F. Rainfall Estimation Based on the Intensity of the Received Signal in a LTE/4G Mobile Terminal by Using a Probabilistic Neural Network. *IEEE Access* **2018**, *6*, 30865–30873. [CrossRef]
23. Jhong, Y.-D.; Chen, C.-S.; Lin, H.-P.; Chen, S.-T. Physical Hybrid Neural Network Model to Forecast Typhoon Floods. *Water* **2018**, *10*, 632. [CrossRef]
24. Alam, K.M.R.; Siddique, N.; Adeli, H. A Dynamic Ensemble Learning Algorithm for Neural Networks. *Neural Comput. Appl.* **2020**, *32*, 8675–8690. [CrossRef]
25. Zhou, J.; Peng, T.; Zhang, C.; Sun, N. Data Pre-Analysis and Ensemble of Various Artificial Neural Networks for Monthly Streamflow Forecasting. *Water* **2018**, *10*, 628. [CrossRef]
26. Haidar, A.; Verma, B.; Sinha, T. A Novel Approach for Optimizing Ensemble Components in Rainfall Prediction. In Proceedings of the 2018 IEEE Congress on Evolutionary Computation (CEC), Rio de Janeiro, Brazil, 8–13 July 2018; pp. 1–8.
27. Kim, T.; Shin, J.; Kim, H.; Heo, J. Ensemble-Based Neural Network Modeling for Hydrologic Forecasts: Addressing Uncertainty in the Model Structure and Input Variable Selection. *Water Resour. Res.* **2020**, *56*, e2019WR026262. [CrossRef]
28. Granata, F.; Di Nunno, F. Forecasting Evapotranspiration in Different Climates Using Ensembles of Recurrent Neural Networks. *Agric. Water Manag.* **2021**, *255*, 107040. [CrossRef]
29. Althoff, D.; Rodrigues, L.N.; Bazame, H.C. Uncertainty Quantification for Hydrological Models Based on Neural Networks: The Dropout Ensemble. *Stoch. Environ. Res. Risk Assess.* **2021**, *35*, 1051–1067. [CrossRef]
30. Bandyopadhyay, G.; Chattopadhyay, S. Single Hidden Layer Artificial Neural Network Models versus Multiple Linear Regression Model in Forecasting the Time Series of Total Ozone. *Int. J. Environ. Sci. Technol.* **2007**, *4*, 141–149. [CrossRef]
31. Zhong, K.; Song, Z.; Jain, P.; Bartlett, P.L.; Dhillon, I.S. Recovery Guarantees for One-Hidden-Layer Neural Networks. *arXiv* **2017**, arXiv:1706.03175.
32. Moriasi, D.N.; Arnold, J.G.; van Liew, M.W.; Bingner, R.L.; Harmel, R.D.; Veith, T.L. Veith Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. *Trans. ASABE* **2007**, *50*, 885–900. [CrossRef]
33. Lagouvardos, K.; Kotroni, V.; Bezes, A.; Koletsis, I.; Kopania, T.; Lykoudis, S.; Mazarakis, N.; Papagiannaki, K.; Vougioukas, S. The Automatic Weather Stations NOANN Network of the National Observatory of Athens: Operation and Database. *Geosci. Data J.* **2017**, *4*, 4–16. [CrossRef]
34. Shen, S.-L.; Zhang, N.; Zhou, A.; Yin, Z.-Y. Enhancement of Neural Networks with an Alternative Activation Function TanhLU. *Expert Syst. Appl.* **2022**, *199*, 117181. [CrossRef]