water

MDPI

Article

Applying Cascade-Correlation Neural Networks to In-Fill Gaps in Mediterranean Daily Flow Data Series

Cristina Vega-Garcia [1,*], Mathieu Decuyper [2,3] and Jorge Alcázar [2]

[1] Department of Agricultural and Forest Engineering, University of Lleida, 25198 Lleida, Spain
[2] Department of Environment and Soil Science, E.T.S.E. Agraria, University of Lleida, 25198 Lleida, Spain
[3] Geo-Information and Remote Sensing Group, Wageningen University and Research, 6708 PB Wageningen, The Netherlands
* Correspondence: cvega@eagrof.udl.es; Tel.: +34-973-702-546

Received: 26 June 2019; Accepted: 10 August 2019; Published: 15 August 2019

Abstract: The analyses of water resources availability and impacts are based on the study over time of meteorological and hydrological data trends. In order to perform those analyses properly, long records of continuous and reliable data are needed, but they are seldom available. Lack of records as in gaps or discontinuities in data series and quality issues are two of the main problems more often found in databases used for climate studies and water resources management. Flow data series from gauging stations are not an exception. Over the last 20 years, forecasting models based on artificial neural networks (ANNs) have been increasingly applied in many fields of natural resources, including hydrology. This paper discusses results obtained on the application of cascade-correlation ANN models to predict daily water flow using Julian day and rainfall data provided by nearby weather stations in the Ebro river watershed (Northeast Spain). Five unaltered gauging stations showing a rainfall-dominated hydrological regime were selected for the study. Daily flow and weather data series covered 30 years to encompass the high variability of Mediterranean environments. Models were then applied to the in-filling of existing gaps under different conditions related to the characteristics of the gaps (6 scenarios). Results showed that when short periods before and after the gap are considered, this is a useful approach, although no general rule applied to all stations and gaps investigated. Models for low-water-flow periods provided better results (r = 0.76–0.8).

Keywords: artificial neural networks; water flow model; hydrological data; data gaps

1. Introduction

Water resources management is based on the study over time of meteorological and hydrological data trends. In order to evaluate resource availability and possible impacts, long records of continuous and reliable data are needed, but they are seldom available. Lack of records (gaps) or discontinuities in data series and quality issues are two of the main problems more often found in databases used for climate studies and water resources management [1–6], especially in mountain regions with limited meteorological monitoring and abundant precipitation often associated to extreme events [7]. Flow data series from gauging stations are also affected by these problems. Unusual flood events may cause breakdowns and failures in the gauging stations which usually result in gaps in the daily flow data series, for instance. Moreover, even very short gaps may compromise the calculation of statistics and data utility [8].

The problem of gaps in data series may be solved theoretically by completing daily flow records from existing data at gauging stations nearby, either upstream or downstream of the same watercourse (e.g., interpolation techniques), although the election of the donor station may be a critical factor affecting the results [8]. Several methods have been used during the last decade for infilling missing

data, including hydrodynamic modeling, remote sensing, or hydrological regionalization based on catchment geomorphological and meteorological data integration, resulting in general multivariate approaches (e.g., [1,3,9]). Calibrated hydrological models for gap-filling streamflow data may perform adequately when estimating general annual trends [10], but in many cases it is not possible to apply these procedures due to lack of data, or they do not provide the accuracy needed to generate daily flow series that reflect the particular characteristics or specificity of each streamflow regime [6].

Approaches used in the past included physically-based rainfall-runoff models, conceptual models (knowledge-driven), or data-driven models [11]. Data-driven, system-theoretical, or black-box models are purely empirical and do not consider the complex physical laws in the real world, but as they depend only on the information content in the hydrological data, they are usually easier to develop [12]. Techniques applied to streamflow prediction or in-filling missing data encompass a great variety of statistical or artificial-intelligence procedures, linear and nonlinear: empirical regression, time series analysis, partitioning modeling, fuzzy rule-based systems, k-nearest neighbor algorithm techniques, pattern recognition, and artificial neural networks (e.g., [9,11,13–20]).

Artificial neural networks (ANN) have proven their value in many complex hydrological modeling problems [6,21–23], often improving results in comparison with other techniques [17,20,24–29]. By using historical data, it is possible to fit the ANN models to the patterns in the data [23,30–32]. They provide many advantages in the robust modeling of nonlinear systems [21,23], but Wu and Chau [20] have pointed out that conclusions in the literature are very inconsistent. Some hybrid models that combine mechanistic and ANN models have also been proposed [12,33] to jointly account for linear and nonlinear trends [34] or to aid with the configuration of ANN models (selection of inputs and/or outputs, e.g., correlograms for Joshi and Patel [27], chaos theory in Elshorbagy et al. [17], moving averages in Wu and Chau [20] and Kashei and Bijari [35]).

Problems most commonly found when working with ANNs are linked to the fact that most authors use variants of the back-propagation model, which architecture is set by trial and error [17,27,36], running many models because they are sensitive to initial weights and have local minima issues. De Vos and Rientjes [11] and Solaimani [12] have dealt with modeling constraints and design aspects that affect model results and performance, and Kalteh [36] has proposed useful approaches to understand the inner representations embedded in the net architecture.

In our study we intended to test the potential of a different ANN type, the cascade-correlation algorithm developed by Fahlman and Lebiere [37] for completing discontinuous daily water flow records in a Mediterranean watershed, coupled with data preprocessing and a genetic algorithm [38] for optimal selection of inputs, an approach used before by Alcazar et al. [21] in environmental flow prediction. The advantage of this model is that its architecture needs not to be set by trial and error like in back-propagation, but optimal hidden layer dimension and net architecture are optimized during the training phase.

Our filling procedure was based on readily available variables, namely, Julian day and precipitation data from existing weather stations located within the watershed area, the usual most important variables to make streamflow predictions [2,23]. Our method could be classified as a single series approach according to the systematization done by Elshorbagy et al. [16], where only one time series is available for the analysis, instead of using two correlated time series (bi-series approach). Nevertheless, an attempt at the identification of Elshorbagy et al.'s groups and modeling the intra-group structure was done through the use of data scenarios. It was expected that the high-variability characteristic of Mediterranean streams would pose more challenges to in-filling than other more stable environments and required considering changing intra (seasonal) and inter-annual flow conditions.

2. Materials and Methods

2.1. Study Area

The domain of study was the Ebro river basin in Spain which is located in the northeast of the country (Figure 1) and has an approximate area of 85,550 km^2 [21]. It has a total length of 910 km, and it is the most important river in Spain in terms of flow, with an average water discharge of 430 m^3/s. The Ebro river is of major importance for ecological and human purposes, being subject to substantial demands from hydropower generation, irrigation of agricultural fields, and recreation and urban uses [1].

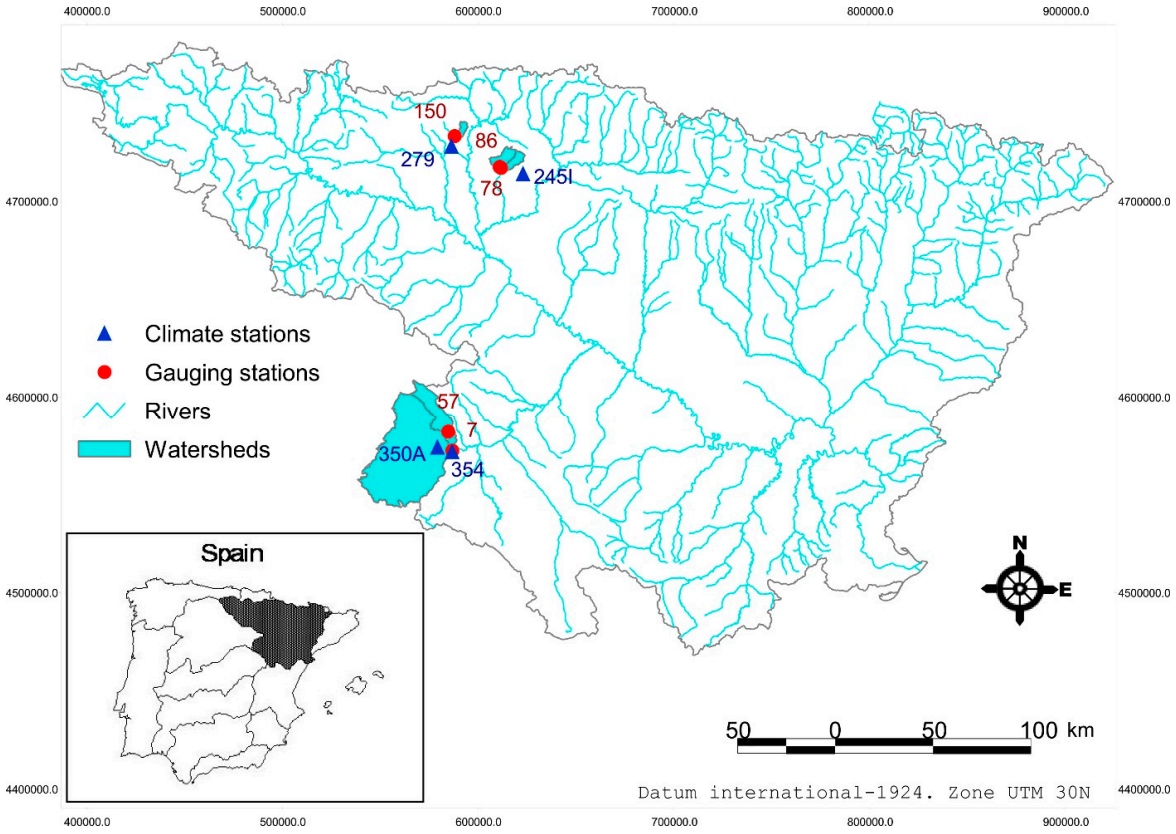

Figure 1. Location of the gauging stations, watersheds, and climate stations used for this study at the Ebro river basin (Spain).

2.2. Database

Five gauging stations (GS) out of the 240 existing within the watershed were selected for this study (Figure 1, Table 1). They were selected because all of them presented unimpaired, natural flow regimes with a reliable data range of 30 years (1976–2005) of daily weather and flow records (average daily water discharge in m^3/s), and no more than three gaps. In the context of this study, we define a gap as a discontinuity in the daily flow data series due to missing data of at least one or more consecutive days. The continuous period of years used for the study was selected based on data reliability and availability, but also to account for the variability of the Mediterranean climate in this area. The study period included dry spells in the 1980s and 1990s, but also periods in the 1970s and 2000s that were humid in Spain. The selected gauging stations showed a mainly rainfall-dominated hydrological regime. Table 1 and Figure 1 show the location and main characteristics of the weather and gauging stations.

Table 1. Location and main characteristics of weather and gauging stations.

Gauging Stations								
Code	**River**	**City**	**Coord. UTM (Zone 30N)**		**GS Altitude (m.o.s.l.)** [1]	**Max. Altitude (m.o.s.l.)** [1]	**GS Watershed Area (km²)**	**GS Mean Day Flow (m³/s)**
			X	**Y**				
7	Jalón	Cetina	586,579	4,572,511	670	1370	1600	0.20
57	Deza	Embid de A.	585,145	4,582,454	782	1309	207	0.17
78	Cemborain	Garinoain	611,419	4,717,013	490	1030	47	0.30
86	Zidacos	Barasoaín	610,830	4,717,489	494	990	76	0.22
150	Salado	Estenoz	588,002	4,733,510	480	1009	28	0.18
Weather Stations								
Code		**City**	**Coord. UTM (Zone 30N)**		**Altitude (m.o.s.l.)** [1]		**Corresponding Gaug. Stations**	
			X	**Y**				
9245I		Lerga	622,963	4,714,022	615		78, 86	
9279		Alloz (Reser.)	586,299	4,728,155	475		150	
9350A		Ariza	579,125	4,574,114	700		7, 57	
9354		Cetina	586,710	4,571,825	680		7, 57	

[1] Meters over sea level.

The GS databases presented a total of 11 gaps (Table 2) in flow records ranging from 6 to 272 days (nine months) out of a possible maximum of 11,950 records (30 years × 365 days). Missing daily flow data values for each gap would be modeled from meteorological data provided by four nearby weather stations. The variables used for completing gaps in the database were Julian day (JD), the precipitation (mm) on the day we had the missing value (PP), and the precipitation on the five days before the date of the missing value (PP-1 to PP-5). The calculated time of concentration using Kirpich's equation [39] for the different watersheds ranged from 1 to 9 h. Furthermore, from the analysis of the precipitation and streamflow data series, we found that the response of watersheds to rainfall events translates in the streamflow data series as peak flows always within a period of 5 days after the weather event. So, given the size of the watersheds and giving time enough for water from the watershed divide to reach the watershed outlet (time of concentration), a period of 5 days was considered appropriate for the particularities of shape, topography, vegetation, and soil characteristics of the watersheds studied.

Table 2. Dates and dimensions of the gaps for each gauging station, type of year and season.

GS	Date and Size of Gaps	Type of Year	Season
7	30 June–23 August 1984 (55 days)	normal	low water flow
7	16 February–12 May 1988(87 days)	wet	high water flow
7	18 June–4 July 1991 (17 days)	dry	low water flow
57	11–30 June 1991 (20 days)	dry	low water flow
78	17 November–14 December 1992 (28 days)	normal	medium water flow
78	8 March–19 July 2004 (134 days)	normal	high-low
86	1 October 1992–29 June 1993 (272 days)	normal	low-high-low
86	13–18 July 2005 (6 days)	normal	low water flow
150	26 July–12 September 1986 (49 days)	normal	low water flow
150	24 February–11 May 1994 (77 days)	normal	high water flow
150	16 October–16 November 2004 (32 days)	normal	low water flow

2.3. ANN Models

Besides the fact that ANN models have proved highly accurate in many previous hydrological applications [6,21–23], in this case, the use of ANN had additional advantages since weather variables would be expected to be highly correlated spatially and temporally, violating assumptions required for traditional statistical model building [21].

There are many types of artificial neural networks, but a specific multilayered feed-forward type of network was used in this study, the cascade-correlation model defined by Fahlman and Lebiere [37] (CCANN). This algorithm was successfully used by Alcázar et al. [21] for the estimation of environmental flows, and a similar model building procedure was followed here. This algorithm has

the advantage of optimizing network architecture in the so-called 'training' or 'learning' process, so it does not rely on trial-and-error for final architecture like back-propagation.

Any net model was initially built with an input layer (with nodes for JD and PP variables) and an output layer (with one output node for mean daily flow), and no nodes in the hidden layer. Learning proceeded by testing nodes for the hidden layer, altering their weights iteratively, and adding these new nodes when they produced an improved net performance. We tested up to three new nodes (instead of just one) for addition at each step of the iterative process, until no improvement in performance was gained. An important difference with the previous environmental flow work (with only 46 training cases) was that in the current gaps problem the number of cases was on the order of magnitude of the thousands, so they did not limit network architecture and size. No constraints were imposed for restricting the number of weights in the model, though this algorithm was designed for optimal structure, and resulting models are usually parsimonious. As the number of weights (degrees of freedom) was not an issue, nodes of our models were always fully connected by weights in three layers (input-to-hidden, hidden-to-output, input-to-output).

We randomly split the databases for analysis in training (56%), test (24%), and validation (20%) groups, as usual in neural network modeling (e.g., [12,27]). To avoid effects of this type of hold-out method of cross-validation in model performance and testing [40], the random splitting was repeated at least three times for each model, and 5 initial replicas were built with different sets of random weights for each group at the beginning of training. Convergence of the 15 trials for each model to a same or similar structure was considered a trait of robustness of the solution.

The iterative learning algorithm was based on an adaptative gradient learning rule [37,41], a variant of the general algorithm of back-propagation [42,43]. Training performance was set to optimize the Pearson product-moment correlation (r, Equation (1)) between observed and predicted outcomes: the known flow values (average daily water discharge in m^3/s) and the output of the net. The test dataset was used to prevent overtraining; learning from the training set was periodically stopped to compute r for the test set and did not continue updating weights once test r started to decrease and diverge from training r. Once training was concluded, the validation dataset was run through the network and its r computed. Predictive model performance was evaluated based on balanced r's for the three datasets, as r is the best known quantitative measure of performance among the group of measures that preserve the pattern of data [40]. Additional criteria used to evaluate the models were based on model residuals, or differences between observed and predicted mean daily flows, specifically the root mean square error (RMSE), mean absolute error (MAE), and absolute maximum error (AME) (Equations (1)–(4)) [40].

$$\text{Root Mean Square Error (RMSE)} \quad \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2} \tag{1}$$

$$\text{Mean Absolute Error (MAE)} \quad \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \widehat{y_i}\right| \tag{2}$$

$$\text{Absolute Maximum Error (AME)} \quad max\left|y_i - \widehat{y_i}\right| \tag{3}$$

$$r = \frac{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)\left(\widehat{y_i} - \overline{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}\sqrt{\sum_{i=1}^{n}\left(\widehat{y_i} - \overline{y}\right)^2}} \tag{4}$$

where $\overline{y}$ is the mean of observed values (y_i observations) and $\widehat{y_i}$ are the predicted values.

Several transformations were applied to the independent variables and tested through a genetic algorithm (GA, [38,44]) provided by Predict® 3.24 software [45] to determine the optimal inputs to the models, previously to model building. In this case, the GA evolved a population of variable sets

that were selected, mated, and mutated for a maximum number of 50 generations (patience = 7) in order to find the fittest combination of independent variables for each network model. Reproduction parameters applied were: cross-over probability = 0.7, mutation = 1.0, and elitist fraction = 0.05. Fitness was evaluated by a logistic multiple regression function for each individual (or subset of the model's input variables, between 2–30 transformed variables) in the population. Once a best model was obtained, a sensitivity analysis through partial derivatives [45] of the variables included as inputs was used to rate their relative importance in the models [21].

2.4. Scenarios

Six different approaches were defined to optimize the predictive models of daily water flow for all and each gap in any gauging station (Table 3). They were required because the natural variability of flow regime (inter and intra-annual variability) is a key aspect in defining the functioning and structure of a river (e.g., [46–49]), especially in Mediterranean ecosystems. However, this high variability of the streamflow database could induce large errors when developing prediction models from large time spans (30 years, Scenario 1). Consequently, several scenarios were proposed to reduce errors induced by a high inter- and intra-annual daily flow variability while keeping the observations for model building representative (Scenarios 2, 3, 4, 6).

We also considered that the hydrological response of watersheds is influenced by the basin characteristics that regulate runoff, such as geomorphology, geology, and vegetation cover. Changes in land cover over time due to either natural causes or human activities may vary the hydrological behavior of the watershed, i.e., the relationship between precipitation and runoff. Then, the accuracy of streamflow prediction models based only on precipitation data over a long period of time may be lowered by errors related to vegetation cover changes, for instance. Scenario 5 tried to reduce this possible source of error by reducing the time span of the observations used for modeling to the short-term.

Table 3. Definition of the scenarios considered in the predictive models.

Scenarios	Conditions
Scenario 1	General model, the whole range of 30 years of weather and flow data was used for all gaps independently of year or season.
Scenario 2	Extreme years (outliers), if present, were removed, only when the gap was not located within those years.
Scenario 3	Inter-annual variability of flow regimes was considered. All available years were classified into three types, wet[1], normal[2], and dry[3] year, based on the characterization of the regime's inter-annual variability, and according to the following criteria [50]. Only years of the same type as those where the gaps were occurring were used for the corresponding analysis.
Scenario 4	Intra-annual variability of flow regimes was considered. Annual flow regimes were divided in low, medium, and high flow periods based on an analysis of seasonal flow variability (similar criteria than in scenario 3 was followed but applied to average monthly values instead of annual values), and only data of the same seasonal period as the gaps were used.
Scenario 5	Short periods of time we selected where the basin characteristics could be considered invariable and therefore the hydrological response of the watershed did not change. Only data of the two years before and after the gap were used for the analysis.
Scenario 6	Combined scenarios 4 and 5. Only data of the same seasonal flow period of the two years before and after the gap were used for the analysis.

[1] A year was considered to be wet if its annual volume in natural regime was greater than the volume corresponding to the 25% exceedance percentile. [2] A year was considered to be normal/average if its annual volume in natural regime laid between the volume corresponding to 25% and 75% exceedance percentile. [3] A year was considered to be dry if its annual volume in a natural regime was lower than the 196 volume corresponding to the 75% exceedance percentile.

Consequently, we built models for each GA (5 GS models), for each gap in any GS (11 gap models), and for these six scenarios, which raised the number of models built, validated, and analyzed for best performance to 54, with 15 replicas each.

3. Results

Best models for each gauging station (GS), scenario, and gap are presented in Tables 4–8, if the models were judged robust and adequate according to the performance criteria (r and RMSE in tables, AME and MAE data not shown).

The most important results for the three gaps in GS-7 *Jalón-Cetina* can be examined in Table 4. The results of GS-7 were based on climate stations 354 and 350A, but in some models one whole station's variables were left out. Models for scenarios S1, S2, and S3 all had $r < 0.50$, with S3 < 0.2. The scenario S2 was the best for the longer 88' spring gap (87 days), but performance was not very good in the validation dataset, probably for the difficulty in modeling variable high water flows. The seasonal scenario S4 was best for the small 1991 summer gap (17 days). Best results were achieved for the 55-days gap in summer 1984 in which the short-term S5 model reached $r = 0.76$ for the training dataset and $r = 0.82$ for the test and validation datasets.

Table 4. Best model results for GS-7 (three gaps). Acronyms: GS (gauging station); CS (climate station); S (scenario #); *r* (Pearson's *r* correlation coefficient, $p < 0.05$); RMSE (root mean square error).

GS	CS	S	Training *r*/*RMSE*	Test *r*/*RMSE*	Validation *r*/*RMSE*	Date and Size of Gaps
7	354 and 350A	5	0.761 *0.257*	0.819 *0.184*	0.817 *0.280*	1984 (55 days)
7	354 and 350A	2	0.734 *0.367*	0.620 *0.486*	0.522 *0.441*	1988 (87 days)
7	354	4	0.704 *0.418*	0.700 *0.444*	0.619 *0.470*	1991 (17 days)

Station GS-57 *Deza-Embid de Ariza* was nearby and we used the same climate stations as for GS-7. Analysis of the water flow curves over all the years based on the average monthly volume showed no extreme years. No homogenous period within the year could be identified (heterogeneous data across all months), so scenarios S2, S3, S4, and S6 could not be applied. The other scenarios gave unsatisfying correlation values $r < 0.50$ for all replicas. Best results can be seen in Table 5.

Table 5. Best model results for GS-57 (one gap).

GS	CS	S	Training *r*/*RMSE*	Test *r*/*RMSE*	Validation *r*/*RMSE*	Date and Size of Gaps
57	354 and 350A	5	0.558 *0.068*	0.406 *0.085*	0.393 *0.103*	1991 (20 days)

Station GS-78 *Cemborain-Garinoain* water flow did not show any extreme years. The two existing gaps overlapped the periods with homogenous very low water flow and the period with high heterogeneous water flow so also here S2, S3, and S6 could not be applied. S1 and S4 gave low r values. The S5 model produced the best results for the 28-day 1992 winter gap with an $r = 0.60$ for the trained dataset and an $r = 0.69$ for the validation dataset (Table 6). The 2004 spring-summer gap (134 days) was not successfully modeled under any of the scenarios.

Table 6. Model results for GS-78 (two gaps).

GS	CS	S	Training *r*/*RMSE*	Test *r*/*RMSE*	Validation *r*/*RMSE*	Date and Size of Gaps
78	245-I	5	0.602 *0.995*	0.547 *0.642*	0.696 *0.533*	1992 (28 days)

GS-86 *Zidacos-Barasoaínis* shared climate station with GS-78. In this gauging station, all scenarios could be applied for modeling but none of them gave good results (r values ≤ 0.50). The best model results can be seen in Table 7, under S1.

Table 7. Model results for GS-86 (two gaps).

GS	CS	S	Training *r*/*RMSE*	Test *r*/*RMSE*	Validation *r*/*RMSE*	Date and Size of Gaps
86	245-I	1	0.508 *0.375*	0.428 *0.445*	0.444 *0.426*	1992 (272 days) + 2005 (6 days)

GS-150 *Salado-Estenozis* was modeled with independent weather variables from climate station 279. S1 showed r values close to 0.56 for training, test, and validation data for all three gaps. S2, S3, and S6 could not be applied to the longer gaps (49-days and 77-days) because the gaps overlapped both high and low flow periods. Models for the gap in the high period (32-days) gave bad correlations in general (r values ≤ 0.50), except for S5. S5 provided reasonably good results for all the gaps (Table 8).

Table 8. Model results for GS-150 (three gaps).

GS	CS	S	Training *r*/*RMSE*	Test *r*/*RMSE*	Validation *r*/*RMSE*	Date and Size of Gaps
150	279	5	0.6625 *0.156*	0.662 *0.153*	0.610 *0.203*	1986 (49 days)
150	279	5	0.7395 *0.112*	0.758 *0.081*	0.717 *0.167*	1994 (77 days)
150	279	5	0.711 *0.173*	0.772 *0.129*	0.809 *0.116*	2004 (32 days)

In most models, and in all best models, architectures were parsimonious and solutions converged to similar nets (Table 9). The differences between the r values of the training, test, and validation datasets were well balanced (similar values for training, test, and validation groups) indicating good reliability in the best models, and RMSEs were low (metric in the same units as the flow data).

The sensitivity analysis of the variables in the best models allowed to identify the most relevant in predicting daily flows. Variables excluded from most models or with partial derivatives that did not indicate relevant contribution were usually the precipitation values 4 and 5 days before the gaps, but there were no clear trends across gauging stations. PP-3 was present in many models as quite influential. Julian day was always in the models, but usually not as the most influential variable.

Table 9. Main characteristics of the models selected for the different gaps.

GS	Date and Size of Gaps	S	ANN Structure	Most Influential Independent Variables
7	1984 (55 days)	5	12-5-1	PP-1 (354), PP-1 (350a)
7	1988 (87 days)	2	13-14-1	PP-4 (350A), PP-3, PP-2 (354)
7	1991 (17 days)	4	8-7-1	PP-5, PP-3, PP, PP-2 (354)
57	1991 (20 days)	5	12-12-1	PP3, PP-5 (354)
78	1992 (28 days)	5	13-4-1	PP, PP-4 (245-I), JD
86	1992 (272 days) + 2005 (6 days)	1	6-12-1	PP-3 (245-I), JD
150	1986 (49 days)	5	9-11-1	PP-3, PP5 (279)
150	1994 (77 days)	5	6-10-1	PP-1 (279)
150	2004 (32 days)	5	6-7-1	PP-3, PP-2, JD (279)

4. Discussion

Natural and rainfall-dominated flow regime watersheds with a reliable data range of 30 years of daily weather and flow records were not abundant in the Ebro river watershed. These conditions were not easily met and consequently, the data used in this study was limited to five gauging stations and the period 1976–2005. Nevertheless, the selected gauging stations were located north and south of the

main stream and include paired, nearby gauging stations. Models to fill eleven gaps representative of different conditions of inter and intra-annual flow variability were built.

The variables used to build the models were purposely few, because Julian day and precipitation data are considered the most important variables to make predictions on streamflow [2,23,51] and they are usually available or easy to gather. Data availability is always an issue in this type of studies. Existing data at gauging stations nearby, either upstream or downstream of the same watercourse, are rarely available. Complex rainfall-runoff models can be built but they also require an abundance of data, and watershed characteristics (such as soil and vegetation influence on water cycle) are not always easily acquired.

Instead, we tested simple models based on generally available weather data and Julian day. In many instances the cascade-correlation network models further reduced the number of variables in order to improve training results; mainly precipitation values delayed 4 and 5 days from the gaps. This fact backed our assessment of the influence of the relatively small size of most of the watersheds and our selection of precipitation variables (previous 1–5 days) based on time of concentration and trends in the data flow series.

Like in other hydrological problems [6,21–23,51], ANNs have proven their potential value for modeling complex hydrological processes with limited data (variables), but the variability of the Pearson r correlation values between observed and predicted outcomes under different scenarios and gauging stations indicate that procedures cannot be generalized. Not all CCANN models performed well enough for their intended in-filling gaps application, even with the advantages provided by this algorithm compared to backpropagation, and by the GA used for selecting the best combination of inputs as a preprocessing technique.

The results of the study suggested that there was not one single scenario suitable for filling up gaps in all gauging stations, but the short-term S5 (two years before and after the gap) gave the best results. Different scenarios would have to be tested, if applicable, but our approach seemed promising if seasonal variability is accounted for and short periods before and after the gap are considered. Using the full 30 years of data (S1) did not give satisfactory results which probably were related to changes in water flow over the years, with higher variability more likely within longer time spans. Gaps in low water flow periods apparently gave better modeling results, probably caused by a lower variability in the data typical of these periods. High heterogeneity in the water flow data negatively influenced the training of suitable models, like in case of GS-57 and GS-86 gauging stations, where suitable models were not found for any scenario or gap. Future work may have to look into neural network algorithms better suited to identify extreme values instead of general trends.

5. Conclusions

Our purpose was to complete discontinuous streamflow data series using a simple filling procedure based on readily available variables. Cascade-correlation neural network models (CCANNs) were built for the estimation of daily water flow in five gauging stations with rainfall-dominated natural hydrological regime located in watersheds of the Ebro river. Models were based on Julian day and precipitation variables from weather stations nearby. We explored the use of a 30-year database under different conditions related to existing gaps in the five gauging stations. Scenarios were defined in order to analyze the performance of the models in different conditions related to intra- and inter-annual natural variability of flow regime as well as database length and characteristics. We concluded that when seasonal variability is accounted for and short periods before and after the gap are considered, CCANNs models can be a very useful predicting tool for filling gaps in streamflow series. No general rule applied to all stations and gaps investigated, on the contrary, individual models had to be built for individual gauging stations using the most appropriate scenario (database length and characteristics, and variables included in the model) to provide best results for each flow series and gap. Models for low water flow periods apparently performed better, probably because of the lower variability in the data typical of these periods.

Author Contributions: Conceptualization, C.V.-G. and J.A.; Data curation, M.D.; Formal analysis, M.D.; Funding acquisition, J.A.; Methodology, C.V.-G. and J.A.; Supervision, C.V.-G. and J.A.; Writing—original draft, C.V.-G., M.D. and J.A.

Funding: This study was partially funded by a research contract between ENDESA S.A. and the University of Lleida.

Acknowledgments: We thank the Ebro River Basin Authority (Confederación Hidrográfica del Ebro) for the use of water flow data. We gratefully acknowledge comments by three reviewers of Water that helped to improve the manuscript.

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

References

1. Alcázar, J.; Palau, A. Establishing environmental flow regimes in a Mediterranean watershed based on a regional classification. *J. Hydrol.* **2010**, *388*, 41–51. [CrossRef]
2. Cuadrat, J.M.; Saz, M.A.; Vicente-Serrano, S.M.; González-Hidalgo, J.C. Water resources and precipitation trends in Aragon. *Int. J. Water Resour. Dev.* **2007**, *23*, 107–123. [CrossRef]
3. Ekeu-wei, I.T.; Blackburn, G.A.; Pedruco, P. Infilling Missing Data in Hydrology: Solutions Using Satellite Radar Altimetry and Multiple Imputation for Data-Sparse Regions. *Water* **2018**, *10*, 1483. [CrossRef]
4. He, J.; Valeo, C.; Chu, A.; Neumann, N.F. Prediction of event-based stormwater runoff quantity and quality by ANNs developed using PMI-based input selection. *J. Hydrol.* **2011**, *400*, 10–23. [CrossRef]
5. Krueger, T.; Freer, J.; Quinton, J.N.; Macleod, C.J.A.; Bilotta, G.S.; Brazier, R.E.; Butler, P.; Haygarth, P.M. Ensemble evaluation of hydrological model hypotheses. *Water Resour. Res.* **2010**, *46*, W07516. [CrossRef]
6. Poff, N.L.; Tokar, S.; Johnson, P. Stream hydrological and ecological responses to climate change assessed with an artificial neural network. *Limnol. Oceanogr.* **1996**, *41*, 857–863. [CrossRef]
7. Rood, S.B.; Pan, J.; Gill, K.M.; Franks, C.G.; Samuelson, G.M.; Shepherd, A. Declining summer flows of Rocky Mountain rivers: Changing seasonal hydrology and probable impacts on floodplain forests. *J. Hydrol.* **2008**, *349*, 397–410. [CrossRef]
8. Harvey, C.L.; Dixon, H.; Hannaford, J. Developing best practice for infilling daily river flow data. Role of Hydrology in Managing Consequences of a Changing Global Environment. In Proceeding of the BHS 3rd International Symposium, Newcastle, UK, 19–23 July 2010; Kirby, C., Ed.; British Hydrological Society: Wallingford, UK; pp. 816–823.
9. Ben Aissia, M.A.; Chebana, F.; Ouarda, T.B.M.J. Multivariate missing data in hydrology—Review and applications. *Adv. Water Resour.* **2017**, *110*, 299–309. [CrossRef]
10. Zhang, Y.; Post, D. How good are hydrological models for gap-filling streamflow data? *Hydrol. Earth Syst. Sci.* **2018**, *22*, 4593–4604. [CrossRef]
11. De Vos, N.J.; Rientjes, T.H.M. Constraints of artificial neural networks for rainfall-runoff modelling trade-offs in hydrological state representation and model evaluation. *Hydrol. Earth Syst. Sci.* **2005**, *9*, 111–126. [CrossRef]
12. Solaimani, K. Rainfall-runoff prediction based on artificial neural network (a case study: Jarahi watershed). *Am.-Euras. J. Agric. Environ. Sci.* **2009**, *5*, 856–865.
13. Abbott, M.B.; Bathurst, J.C.; Cunge, J.A.; O'Connell, P.E.; Rasmussen, J. An introduction to the European hydrological system—Système Hydrologique Européen, SHE, 1: History and philosophy of a physically-based, distributed modelling system. *J. Hydrol.* **1986**, *87*, 45–59. [CrossRef]
14. Abbott, M.B.; Bathurst, J.C.; Cunge, J.A.; O'Connell, P.E.; Rasmussen, J. An introduction to the european hydrological system—Système Hydrologique Européen, SHE, 2: Structure of a physically-based, distributed modelling system. *J. Hydrol.* **1986**, *87*, 61–77. [CrossRef]
15. Burnash, R.J.C. The NWS River Forecast System—Catchment modeling. In *Computer Models of Watershed Hydrology*; Singh, V.P., Ed.; Water Resources Publications: Highlands Ranch, CO, USA, 1995; pp. 311–366.
16. Elshorbagy, A.; Panu, U.S.; Simonovic, S.P. Group-based estimation of missing hydrological data: I. Approach and general methodology. *Hydrol. Sci. J.* **2000**, *45*, 849–866. [CrossRef]
17. Elshorbagy, A.; Simonovic, S.P.; Panu, U.S. Estimation of missing streamflow data using principles of chaos theory. *J. Hydrol.* **2002**, *255*, 123–133. [CrossRef]

18. Gao, Y.; Merz, C.; Lischeid, G.; Schneider, M. A review on missing hydrological data processing. *Environ. Earth Sci.* **2018**, *77*, 47. [CrossRef]
19. Kamwaga, S.; Mulungu, D.M.M.; Valmba, P. Assessment of empirical and regression methods for infilling missing streamflow data in Little Ruaha catchment Tanzania. *Phys. Chem. Earth* **2018**, *106*, 17–28. [CrossRef]
20. Wu, C.L.; Chau, K.W. Data-driven models for monthly streamflow time series prediction. *Eng. Appl. Artif. Intell.* **2010**, *23*, 1350–1367. [CrossRef]
21. Alcázar, J.; Palau, A.; Vega-Garcia, C. A neural net model for environmental flow estimation at the Ebro River Basin, Spain. *J. Hydrol.* **2008**, *349*, 44–55. [CrossRef]
22. Araujo, P.; Astray, G.; Ferrerio-Lage, J.A.; Mejuto, J.C.; Rodríguez-Suarez, J.A.; Soto, B. Multilayer perceptron neural network for flow prediction. *J. Environ. Monit.* **2011**, *13*, 35–41. [CrossRef]
23. Besaw, L.E.; Rizzo, D.M.; Bierman, P.R.; Hackett, W.R. Advances in ungauged streamflow prediction using artificial neural networks. *J. Hydrol.* **2010**, *386*, 27–37. [CrossRef]
24. Elshorbagy, A.; Panu, U.S.; Simonovic, S.P. Group-based estimation of missing hydrological data: II. Application to streamflows. *Hydrol. Sci. J.* **2000**, *45*, 867–880. [CrossRef]
25. Dastorani, M.T.; Moghadamnia, A.; Piri, J.; Rico-Ramirez, M. Application of ANN and ANFIS models for reconstructing missing flow data. *Environ. Monit. Assess.* **2009**, *166*, 421–434. [CrossRef] [PubMed]
26. Hsu, K.L.; Gupta, H.V.; Sorooshian, S. Artificial neural network modeling of the rainfall-runoff process. *Water Resour. Res.* **1995**, *31*, 2517–2530. [CrossRef]
27. Joshi, J.; Patel, V.M. Rainfall-runoff modeling using artificial neural network (a literature review). In Proceedings of the National Conference on Recent Trends in Engineering & Technology, Anand, Gujarat, India, 13–14 May 2011.
28. Khalil, M.; Panu, U.S.; Lennox, W.C. Groups and neural networks based streamflow data infilling procedures. *J. Hydrol.* **2001**, *241*, 153–176. [CrossRef]
29. Sajikumar, N.; Thandaveswara, B.S. A non-linear rainfall-runoff model using an artificial neural network. *J. Hydrol.* **1999**, *216*, 32–55. [CrossRef]
30. Dastorani, M.; Moghadamnia, A.; Piri, J.; Rico-Ramirez, M. Application of ANN and ANFIS models for reconstructing missing flow data. *Environ. Monit. Assess.* **2010**, *166*, 421–434. [CrossRef] [PubMed]
31. Dastorani, M.T.; Talebi, A.; Dastorani, M. Using neural networks to predict runoff from ungauged catchments. *Asian J. Appl. Sci.* **2010**, *3*, 399–410. [CrossRef]
32. Kuo, C.C.; Gan, T.Y.; Yu, P.S. Seasonal streamflow prediction by a combined climate-hydrologic system for river basins of Taiwan. *J. Hydrol.* **2010**, *387*, 292–303. [CrossRef]
33. Wu, C.L.; Chau, K.W.; Li, Y.S. Predicting monthly streamflow using data-driven models coupled with data-preprocessing techniques. *Water Resour. Res.* **2009**, *45*, W08432. [CrossRef]
34. Rajurkar, M.P.; Kothyari, U.C.; Chaube, U.C. Artificial neural networks for daily rainfall-runoff modeling. *Hydrol. Sci. J.* **2002**, *47*, 865–877. [CrossRef]
35. Khashei, M.; Bijari, M. An artificial neural network (p, d, q) model for timeseries forescasting. *Expert Syst. Appl.* **2010**, *37*, 479–489. [CrossRef]
36. Kalteh, A.M. Rainfall-runoff modeling using artificial neural networks (ANNs): Modeling and understanding. *Casp. J. Environ. Sci.* **2008**, *6*, 53–58.
37. Fahlman, S.E.; Lebiere, C. *The Cascade-Correlation Learning Architecture, Advances in Neural Information Processing Systems 2*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1990.
38. Koza, J.R. *Genetic Programming*; MIT Press: Cambridge, MA, USA, 1993.
39. Kirpich, Z.P. Time of concentration of small agricultural watersheds. *Civ. Eng.* **1940**, *10*, 362.
40. Bennett, N.D.; Croke, B.F.W.; Guariso, G.; Guillaume, J.H.A.; Hamilton, S.H.; Jakeman, A.J.; Marsili-Libelli, S.; Newham, L.T.H.; Norton, J.P.; Perrin, C.; et al. Characterising performance of environmental models. *Environ. Modell. Softw.* **2013**, *40*, 1–20. [CrossRef]
41. Bridle, J.S. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In *Neuro-computing: Algorithms, Architectures and Applications*; Fougelman-Soulie, F., Herault, J., Eds.; Springer: Berlin, Germany, 1990; pp. 227–236.
42. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Foundations*; Rumelhart, D.E., McClelland, J.L., Eds.; MIT Press: Cambridge, MA, USA, 1986; Volume 1, pp. 318–362.

43. Werbos, P.J. *The Roots of Backpropagation: from Ordered Derivatives to Neural Networks and Political Forecasting*; John Wiley & Sons, Inc.: New York, NY, USA, 1994; 319p.
44. Bowden, G.J.; Dandy, G.C.; Maier, H.R. Input determination for neural network models in water resources applications. Part 1—Background and methodology. *J. Hydrol.* **2005**, *301*, 75–92. [CrossRef]
45. Neuralware. NeuralWorks Predict. In *The Complete Solution for Neural Data Modelling*; NeuralWare: Pittsburg, PA, USA, 2009; 237p.
46. Richter, B.D.; Baumgartner, J.V.; Powell, J.; Braun, D.P. A method for assessing hydrologic alteration within ecosystems. *Conserv. Biol.* **1996**, *10*, 1163–1174. [CrossRef]
47. Poff, N.L.; Allan, J.D.; Bain, M.B.; Karr, J.R.; Prestegaard, K.L.; Richter, B.D.; Sparks, R.E.; Stromberg, J.C. The natural flow regime: A paradigm for river conservation and restoration. *BioScience* **1997**, *47*, 769–784. [CrossRef]
48. Stewardson, M.J.; Gippel, C.J. Incorporating flow variability into environmental flow regimes using the flow events method. *River Res. Appl.* **2003**, *19*, 459–472. [CrossRef]
49. Poff, N.L.; Zimmerman, J.K.H. Ecological responses to altered flow regimes: A literature review to inform the science and management of environmental flows. *Freshw. Biol.* **2010**, *55*, 194–205. [CrossRef]
50. Martínez Santa-María, C.; Fernández Yuste, J.A. IAHRIS 2.2 Indicators of Hydrological Alteration. Methodological Reference Manual. 2010. Available online: https://www.researchgate.net/publication/46387504_Iahris_New_Software_to_Assess_Hydrologic_Alteration (accessed on 28 July 2019).
51. Wu, C.L.; Chau, K.W. Rainfall–runoff modeling using artificial neural network coupled with singular spectrum analysis. *J. Hydrol.* **2011**, *399*, 394–409. [CrossRef]