02-024

RAINFALL-RUNOFF MODELLING AT CATCHMENT SCALE WITH ARTIFICIAL INTELLIGENCE ALGORITHMS

Zubelzu Mínguez, Sergio (1); Matendo, Sara (1); Zanella, Andrea (2); Capuzzo, Martina (2); Bennis, Mehdi (3)

(1) Universidad Politécnica de Madrid, (2) University of Padova, (3) University of Oulu

Black-box type data-driven algorithms provide us with a number of advantages when seeking to model rainfall-runoff processes at catchment scale. The key issue is the correct definition of the system boundaries and the system inputs and outputs as well. Under such configuration, those algorithms allows us to reach great accuracy in modelling hydrological processes overcoming the issues when physically-based models are used. In this paper we address the adequate setting of the abovementioned algoprithms in light of the defined mass conservation law within the catchment. Some initial results are drafted for several catchments and the reliability of the balck-box type approaches is discussed.

Keywords: Hydrology; Rainfall-runoff; Catchment; Machine Learning.

SIMULACIÓN DE PROCESOS RAINFALL-RUNOFF EN CUENCAS MEDIANTE INTELIGENCIA ARTIFICIAL

Los algoritmos de datos tipo caja negra presentan múltiples ventajas para modelizar los procesos rainfall-runoff a escala de cuenca. La clave reside en la correcta delimitación de los límites del sistema y del balance de conservación de la masa dentro de éste. En tales condiciones, esta clase de modelos permiten alcanzar una elevada precisión en la simulación de los procesos de rainfall-runoff mejorando las complicaciones derivadas del uso de modelos de base física. En este trabajo se estudia la correcta definición de modelos de datos a la vista del balance conservación de la masa a la escala de cuenca. Se presentan resultados preliminares para varias cuencas y se discute la validez de los modelos y su enfoque.

Palabras clave: Hidrología; Precipitacìon-escorrentia; Cuenca, Aprendizaje automático

Correspondencia: Sergio Zubelzu. Correo: sergio.zubelzu@upm.es

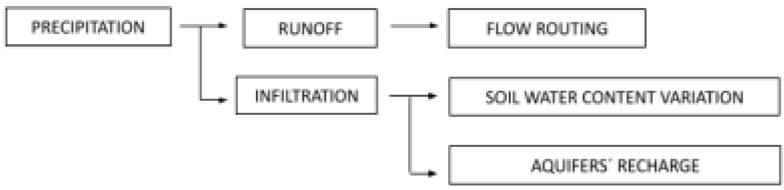
Agradecimientos: La publicación es parte del proyecto PCI2020-120694-2, financiado por MCIN/AEI/10.13039/501100011033 y por la Unión Europea “NextGenerationEU”/PRTR

1. Introduction

Hydrological systems are complex, involving a number of highly variable and interlinked distant processes, from precipitation to flood routing through infiltration, evapotranspiration, soil water movement, among others.

The so-called event-based hydrology (figure 1) deals with storm events and the processes related to.

Figure 1. Event-based hydrological processes flowchart



Our main ambition when dealing with events-based hydrological modelling is to predict the water level evolution at the main river channel during a storm event and after it ceases. From a physical perspective two relevant processes are triggered by the rain: the rainfall infiltration (and consequently the surface runoff) and the flood routing. The infiltration determines the amount, and the rates as well, of water reaching the river channel while the flood routing refers to the temporal distribution of the wave along the water course. Modelling the previous process at catchment scale constitutes a problem of particular complexity since many different processes are involved and a number of difficulties derived from the spatial heterogeneity arise. Physically-based theories presents a number of computational issues, complex parametrization or inability to deal with stochastic process that hinder their use in practical applications. Otherwise, classical empirical models inferred from observations often lack of generality, adaptability and interpretability.

In this context, black-box type data-driven approaches open up interesting opportunities for capturing the complexity of events-based hydrological phenomena to help overcome the aforementioned limitations of classical approaches.

2. Literature review

Machine learning algorithms have received growing interest for modelling hydrological processes. Previous attempts for modelling hydrological processes with data-driven algorithms can be grouped into two categories:

a) Pure autoregressive models for forecasting either isolated hydrological processes or the hydrological system as a whole (see for example Tikhanmarine et al. 2020, Zuo et al. 2020, Shabri and Suhartono 2012, Dariane et al. 2018 or Ebrahini and Sourian 2020).

b) Data-driven models based on physical criteria. Several works move away from autoregressive models and build data-driven models upon apparent physical relationships as for example Mehrpavar and Asghari (2018) or Zhang et al. (2020).

While models framed within the first approach clearly benefit from the ability of data-driven algorithms for capturing hydrological behaviour efficiently, they lack of generality and obviate the physics underlying the phenomena. On the contrary, approaches merging data-driven models and physical approaches provide different opportunities for the accurate modelling of hydrological processes. Suitable data-driven algorithms can be used to extract information from the sensing data and improve the accuracy of process-based hydrologic models, thus offering more reliable predictions of the water fluxes.

Our main ambition in this work is aligned with the previous idea of merging physically-based and AI domains. We seek to analyse the suitability of neural network built following mass conservation law principia for predicting the short-term evolution of rainfall-runoff processes at several catchments located at the Alto Ebro river basin (Cantabria, Spain).

2. Materials and methods

2.1 Case studies

We focus on two different catchments located at the upper basin of the Spanish Ebro river. After a detailed analysis of the available weather information across the area, we identified two catchments suitable to serve as case studies. Both have water level gauges at the downstream-end point what allowed us to define the limits of the system where the mass conservation principia were defined. Figure 2 presents the limits of the studied watersheds.

Figure 2. Catchments

Izarilla-Suso Soto

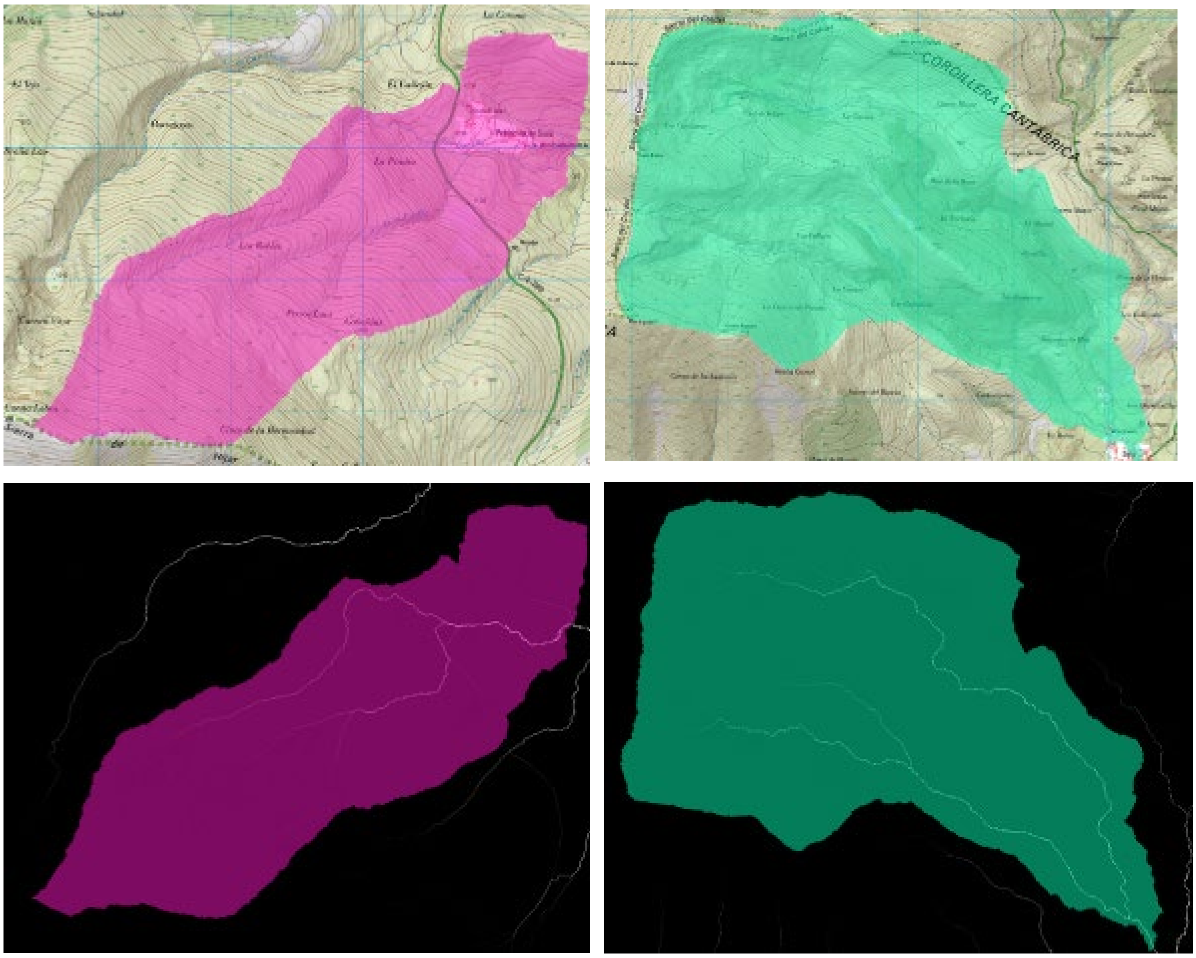


Figure 3 displays a sketch of the water level measurements facilities and a picture of the Soto weather station retrieved from the SAIH EBRO platform.

Figure 3. Sketch of the water level measurement facilities at Izarilla-Suso and a picture of Soto weather station

Izarilla-Suso Soto

From the datasets provided by the Spanish Confederación Hidrográfica del Ebro, we have been able to extract 15-minutes records of water level for the Izarilla-Suso (30/11/2021 00:15 to 11/04/2022 10:00) and Soto (02/12/2021 10:30 to 11/04/2022 10:00) water level gauges and the same records of precipitation from both Soto and Izarilla-Suso weather stations.

2.1 Mass conservation within a catchment

We focus on hydrological catchments defined as portions of land where water inputs, outputs and internal operations can be identified. The transfer function over time (t) for that system can be defined as follows (eq. 1).

$$dS/dt=I(t)-O(t) \tag{1}$$

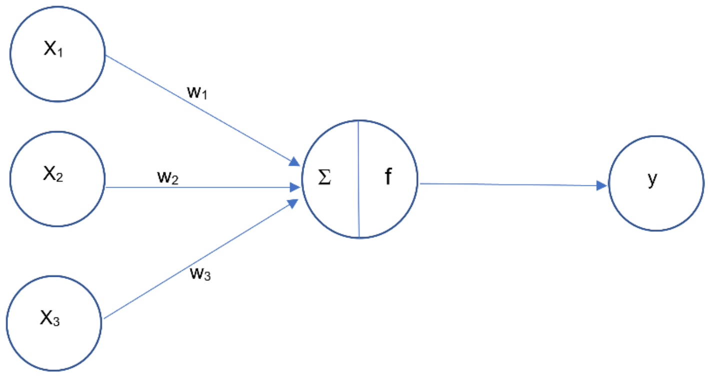
Where I and O are inputs and outputs, respectively, and S stands for internal storages.

2.1 Neural networks

Neural networks can help model the relationship between a set of inputs and outputs signals by mimicking the way biological brains respond to sensory inputs. They are framed within the so-called black-box methods since the mechanisms transforming the inputs in outputs are obfuscated by an imaginary box.

Figure 4 sketches the architecture of a simple neural network.

Figure 4. Neural networks sketch



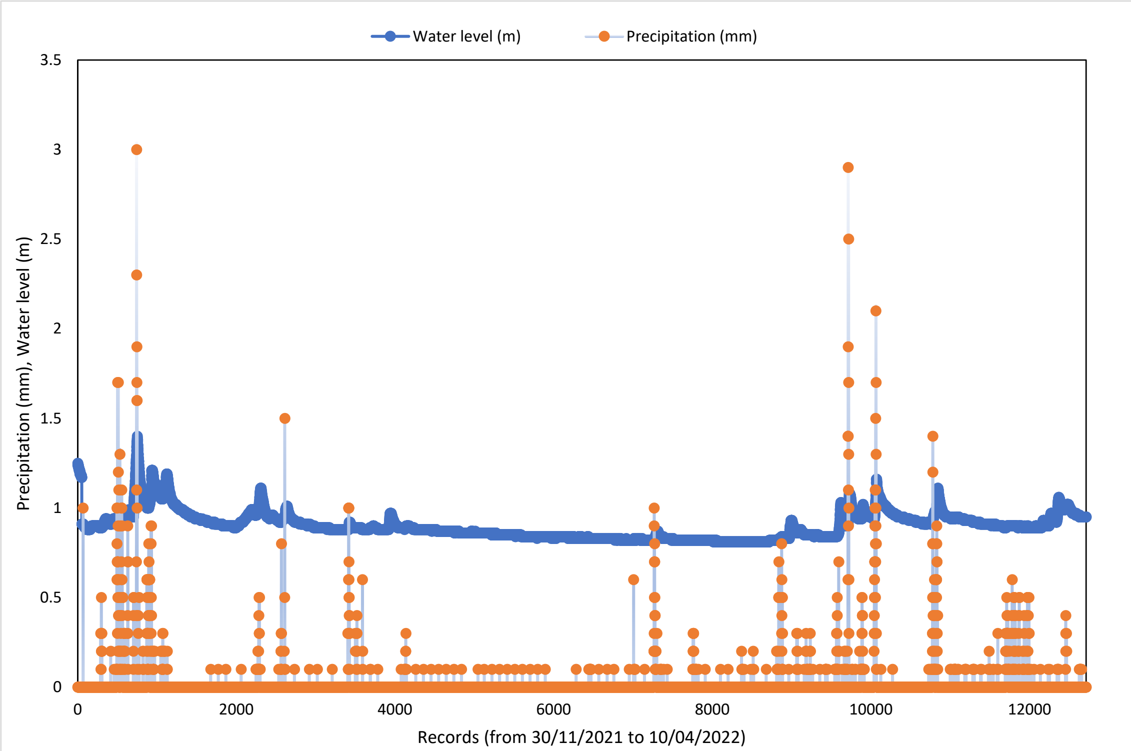
In this work we have used the *train()* function from Caret R package for exploring the best configuration in terms of network topology, activation functions and training algorithm.

3. Results

3.1 Descriptive analysis

As it can be observed from figure 5, the water level kept reasonably constant over time at the analysed period, and it looks that water level increased as a result of storm events.

Figure 5. Precipitation and water level



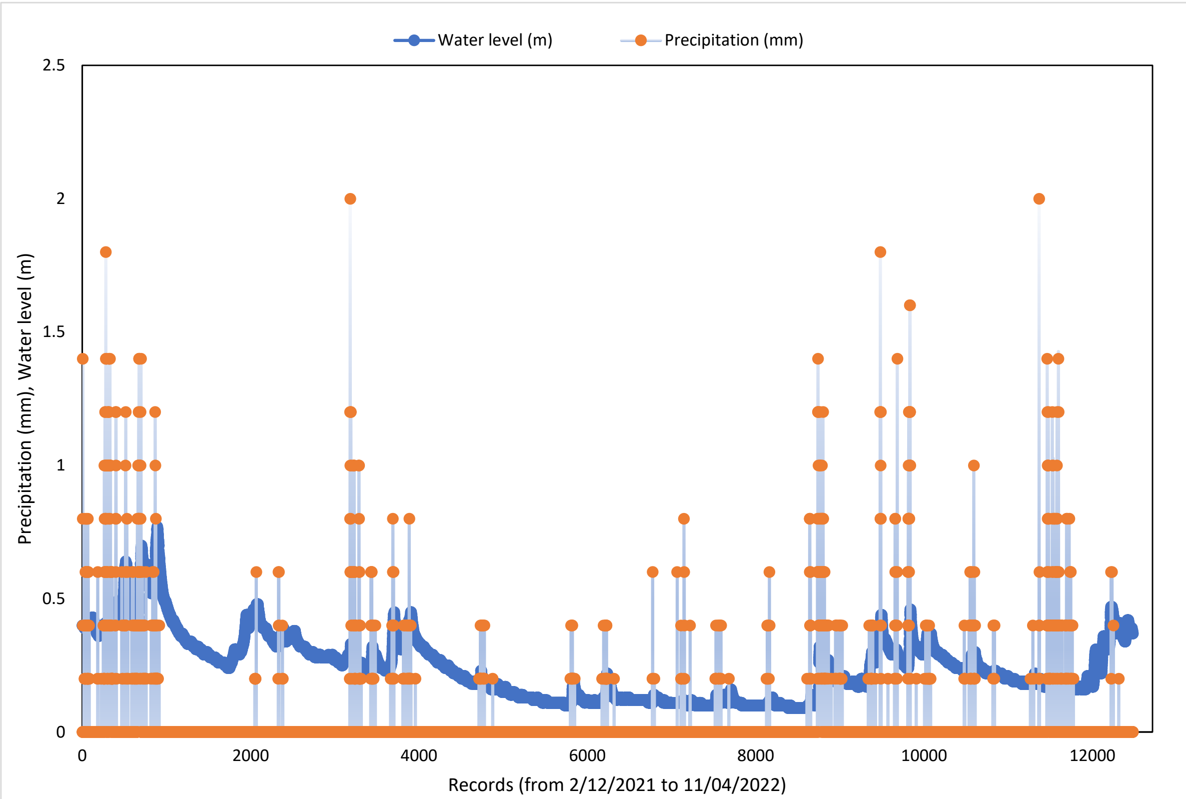


Table 1 presents the main statistics extracted from the data distribution.

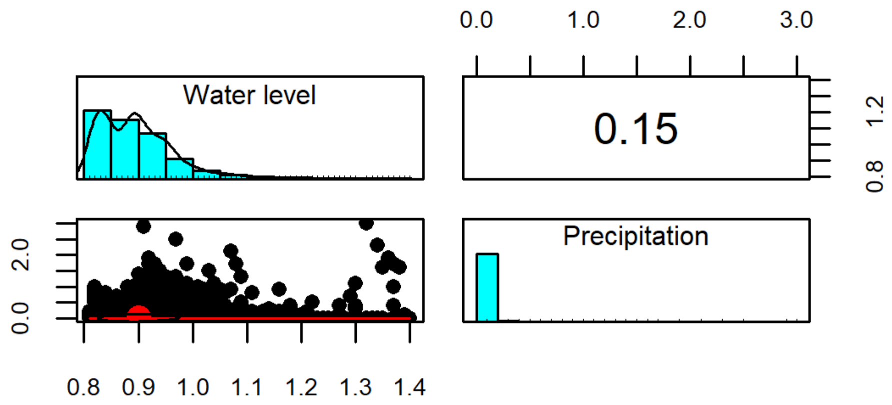
Table 1. Samples statistics

Statistic	Soto		Izarilla-Suso	
	Water level (m)	Precipitation (mm)	Water level (m)	Precipitation (mm)
Mean	0.24	0.04	0.90	0.02
Error	0.00	0.00	0.00	0.00
Median	0.23	0.00	0.89	0.00
Mode	0.11	0.00	0.84	0.00
Standard Deviation	0.12	0.15	0.08	0.12
Variance	0.01	0.02	0.01	0.01
Kurtosis	0.84	36.74	4.46	155.47
Assimetry	0.91	5.54	1.61	10.46
Range	0.77	2.00	0.59	3.00
Minimum	0.00	0.00	0.81	0.00
Maximum	0.77	2.00	1.40	3.00

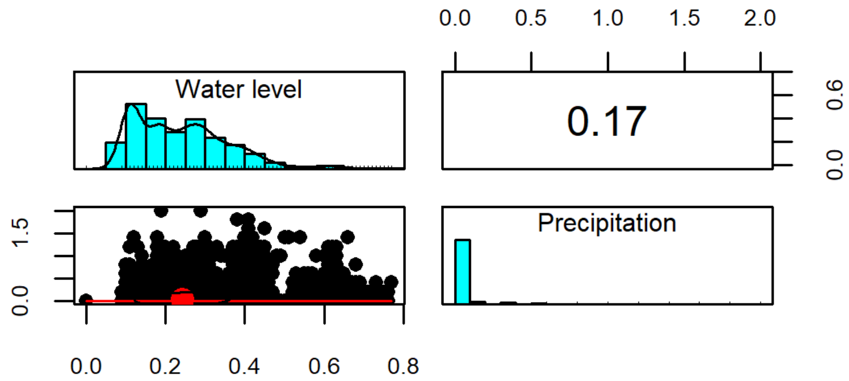
Table 1 shows differences between precipitation records at both weather stations (though they are 15 km far from each other). In general terms, precipitation looks variable, disperse and biased while water levels keep reasonably constant what could point to infiltration or runoff detention smoothing the transference.

From figure 5 it can also be perceived a clear correlation between the storms occurrence and the increase of water level. It looks like storm events make water level to increase (either directly or after a lag time) while dry periods drive to a smoothed water level decrease. To check this assumption we have computed the linear correlation coefficients and the scatterplots (figure 6).

Figure 6. Correlations between water levels and precipitation for Izarilla-Suso (a) and Soto(b) catchments



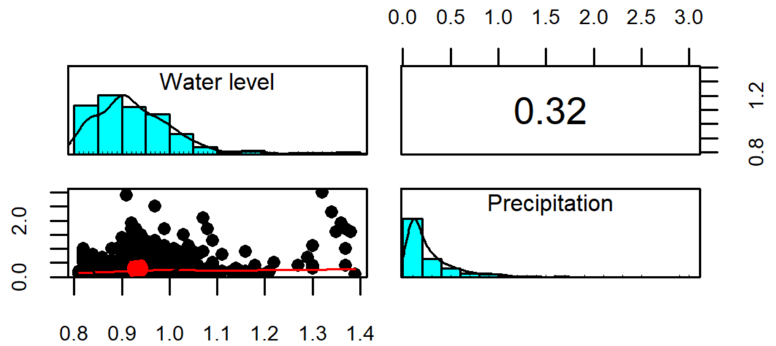
a)



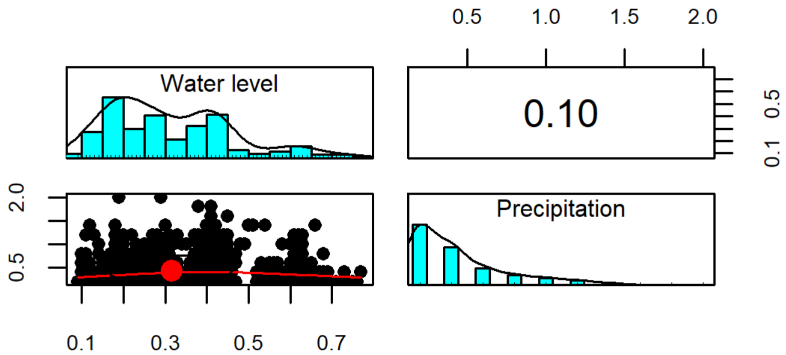
b)

Linear correlations do not look as strong as it could be expected from the preliminary analysis of figure 5. The reality is that a number of zero values of precipitation do not correspond with constant water level records what is altering the global correlation coefficient. If we carefully look at the relationship between non-zero precipitation records and water levels one could expect more intense correlations if the presence of zero precipitation values alter the underlying relationships (see figure 7).

Figure 7. Correlations between non-zero precipitation records and corresponding water levels for Izarilla-Suso (a) and Soto(b) catchments



a)



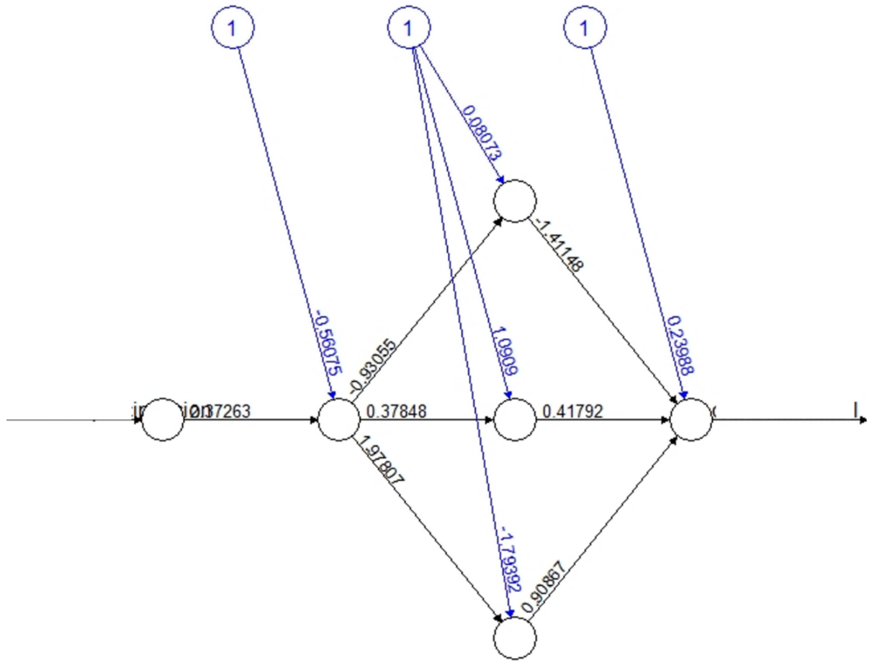
b)

As expected, the linear the correlation figure for Izarilla-Suso increases. On the contrary, the linear correlation coefficient referred to the Soto catchment decreases when non-zero precipitation values are used. Though different reason can cause this behaviour, if we pay attention only to pure hydrological processes, maybe the presence of time dependent factors affecting the transient time can explain this reduction. If this were the reason, the coefficient could increase if we incorporate some lagged data before the storm starts and after it ceases.

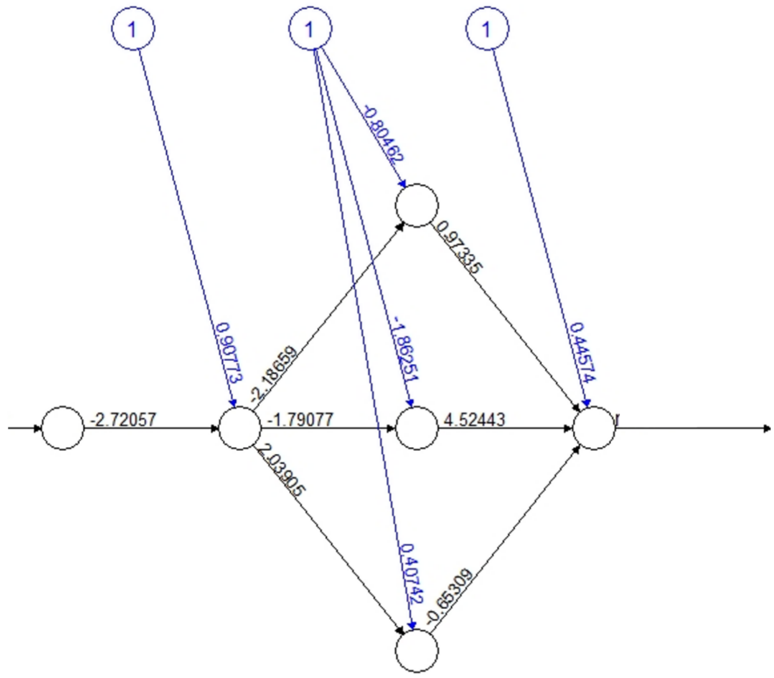
3.2 Water level prediction

As we found previously, the linear correlation coefficients do not report strong correlations between dependent and independent variables. However, it does not mean another type of non-linear relationship could explain the behaviour of water level from precipitation records. Neural networks are particularly suited for capturing complex non-linear relationship between different variables. Upon this idea we have built two ANN models fed with either the entire datasheet or only non-zero precipitation records and corresponding water levels.

Figure 8. Neural networks topology for Izarilla-Suso (a) and Soto (b) catchments



a)



b)

The performance of the best model for Izarilla-suso and Soto are poor since the linear correlation coefficients (between predictions and actual values) for the test data only reached 0.12 and 0.18, respectively. Conversely using the datasheets without the zero precipitation records, the model for Izarilla-Suso performed better reaching a linear correlation coefficient of 0.32 while the model for Soto only reached 0.06.

4. Discussion

Different factors can explain the poor performance of the deduced models and several thoughts arise from that results. First, from a clear mass balance perspective applied to small, simple and relatively homogeneous catchments, we achieved poor results in terms of explainability. From physical and conceptual perspectives, the approach is overwhelming: the unique input (precipitation) has to be transformed into to the unique output (discharge) through the transfer function (affected by detention, infiltration and other internal processes). Similarly, black-box type methods, in particular ANN for such simple case studies, look perfectly suited for modelling the rainfall-runoff process since in those cases, one input is related to one output through a number of highly variable (time and space) interrelated and (supposedly) non linear internal processes. On the other hand, we have used the *train()* function of the *CARET* library using both *neuralnet* (finding the best parameter tuning considering the number of hidden layers) and the *nnet* (finding the best parameter tuning considering jointly the network size – number of hidden units– and the decay) methods. The best model fit is expected to be achieved with those methods since they yield the best a parameter tuning after checking a number of possible combinations.

One could expect that for such simple processes, the proposed approaches should be able to find the best ANN configuration. A number of studies have previously addressed rainfall-runoff modelling by data-driven algorithms. Many reach high accuracy and predictive capacity through very sophisticated algorithms and pre-processing techniques. For example, Ali et al. (2020) developed a complex ensemble model hybridized with both random forest and kernel ridge regression for monthly rainfall forecasting. Farajzadeh and Alizadeh (2003) proposed a mixed method combining ARIMAX and least squares support vector machine. Bui et al., (2020) or Wang et al. (2020) both used neural networks for predicting flood susceptibility areas using different topography and vegetation related variables while Pourghasemi et al. (2020) analysed the suitability of different metaheuristic approaches for flood mapping. Tikhamarine et al. (2019) used artificial neural networks and support vector regression coupled with the Grey wolf optimizer algorithm. Zuo et al. (2020) used different preprocessing techniques and support vector machine. Shabri and Suhartono (2012) comparing predictions from east-squares support vector machine with those obtained with statistical autoregressive integrated moving average or artificial neural network combined with support vector machine using various statistical measures, by Dariane and Azimi (2018) combining neural networks and fuzzy models for monthly forecast or by Ebrahimi and Shourian (2020) for daily streamflow forecasting using a sort of dynamic K-nearest neighbours. Zhang et al (2020) for coming up with an ensemble streamflow prediction derived with bayesian techniques or by Niu et al. (2020) using evolutionary extreme learning machine and variational mode decomposition to predict annual streamflow time series.

All of them achieved relevant predictive capacity but, at the same time, all lack of physical coherence. They used very efficient algorithms and pre-processing techniques to come up with methods seeking to maximize a set of statistical metrics. In most cases they used pure auto-regresive approaches or combined a set of independent variables in a questionable manner (for example feeding a model on both temperature and rainfall to predict streamflow). This approaches can only achieve statistical representativity but they are not suited for predicting in an effective manner the hydrological behaviour.

Other approaches base on coherent conceptual approaches. For example, Alizadeh et al. (2018) used ANN to predict the observed streamflow in a main river using records from several tributary water courses. Similarly, Kratzert et al. (2019) probed that using a simple long short-term memory network for the ensemble modelling of 531 basins achieved better performance than hydrological models that were calibrated individually. These works, more aligned with the idea supported in this manuscript, use the measured streamflow at a set of tributary courses to predict the expected flow at a downstream point. These results point that the approach followed in this work is feasible and that further research must be conducted to ensure the model is better tuned. This further research must be conceived in terms of physical coherence instead of on data processing.

4. Conclusions

Data-driven models can provide useful tools for hydrology always they are clearly shaped by physical criteria. In this work we use artificial neural networks to model the rainfall-runoff processes in two catchments located in the Alto Ebro basin. We did not observe strong relationships between precipitation and water level, what enhances the relevance of internal physical processes determining the different evolution of storm events. Other non-linear relationship could also explain the behaviour.

When we build artificial neural networks for modelling those processes, we have obtained poor results in predicting the test samples probably caused by different factors as for example physical reasons (initial soil water content determining the infiltration rates) or the presence of local water discharges not considered in the model. further analysis should be carried out in relation to the time lags between precipitation occurrence and water flow increase. In our opinion the model performance can be largely improved by going further in the study physical relationships. Only improving the physical coherence of the data-driven approach effective and reliable predictive models can emerge.

5. References

Ali, M., Prasad, R., Xiang, Y., & Yaseen, Z.M. (2020) Complete ensemble empirical mode decomposition hybridized with random forest and kernel ridge regression model for monthly rainfall forecasts. *Journal of Hydrology* 584, 124647, https://doi.org/10.1016/j.jhydrol.2020.124647.

Alizadeh, M.J., Nourani, V., Mousavimehr, M., & Kavianpour, M.R. (2018) Wavelet-IANN model for predicting flow discharge up to several days and months ahead. *Journal of Hydroinformatics*, 20(1), 695 134–148.

Bui, Q.T., Nguyen, Q.H., Nguyen, X.L., Pham, V.D., Nguyen, H.D., & Pham, V.M. (2020) Verification of novel integrations of swarm intelligence algorithms into deep learning neural network for flood susceptibility mapping. *Journal of Hydrology*, 581, 124379, https://doi.org/10.1016/j.jhydrol.2019.124379.

Dariane, A.B., & Azimi, S. (2018) Streamflow forecasting by combining neural networks and fuzzy models using advanced methods of input variable selection. *Journal of Hydroinformatics*, 20(2) 520–532.

Ebrahimi, E., & Shourian, M. (2020) River flow prediction using dynamic method for selecting and prioritizing K-nearest neighbors based on data features. *Journal of Hydrology Engineering*, 25(5), 04020010, https://doi.org/10.1061/(ASCE)HE.1943-5584.0001905.

Farajzadeh, J., & Alizadeh, F.A. (2003) hybrid linear–nonlinear approach to predict the monthly rainfall over the Urmia Lake watershed using wavelet-SARIMAX-LSSVM conjugated model. *Journal of Applied Meteorology and Climatology*, 42 381–388.

Kratzert, F.; Klotz, D.; Shalev, G.; Klambauer, G.; Hochreiter, S. (2019) Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Science*, 23(12), 5089–5110.

Mehrparvar, M., & Asghari, K. (2018) Modular optimized data assimilation and support vector machine for hydrologic modeling. *Journal of Hydroinformatics*, 20(3), 728–738.

Pourghasemi, H.R., Razavi-Termeh, S.V., Kariminejad, N., Hong, H., & Chen, W. F. (2020) An assessment of metaheuristic approaches for flood assessment. *Journal of Hydrology* 582, 124536, https://doi.org/10.1016/j.jhydrol.2019.124536.

Shabri, A., & Suhartono, S. (2012) Streamflow forecasting using least-squares support vector machines. *Hydrological Sciences Journal*, 57(7) 1275–1293.

Tikhamarine, Y., Souag-Gamane, D., Ahmed, A.N.; Kisi, O., & El-Shafie, A. (2020) Improving artificial intelligence models accuracy for monthly streamflow forecasting using greyWolf optimization (GWO) algorithm. *Journal of Hydrology,* 582, 124435, https://doi.org/10.1016/j.jhydrol.2019.124435.

Wang, Y., Fang, Z., Hong, H., & Peng, L. (2020) Flood susceptibility mapping using convolutional neural network frameworks. *Journal of Hydrology*, 124482, https://doi.org/10.1016/j.jhydrol.2019.124482.

Zhang, J., Chen, J., Li, X., Chen, H., Xie, P., & Li, W. (2020) Combining postprocessed ensemble weather forecasts and multiple hydrological models for ensemble streamflow predictions. *Journal of Hydrology Engineering*, 25(1), 04019060, https://doi.org/10.1061/(ASCE)HE.1943-5584.0001871.

Zuo, G., Luo, J., Wang, N., Lian, Y., & He, X. (2020) Two-stage variational mode decomposition and support vector regression for streamflow forecasting. *Hydrology and Earth System Science*, 24(11) 5491–5518.

Niu, W.J., Feng, Z.K., Chen, Y.B., Zhang, H.R., & Cheng, C.T. (2020) Annual streamflow time series prediction using extreme learning machine based on gravitational search algorithm and variational mode decomposition. *Journal of Hydrologic Engineering*, 25(5), 04020008, https://doi.org/10.1061/(ASCE)HE.1943-5584.0001902.

Communication aligned with the Sustainable Development Objectives