Renewable Energy, Environment, and Power Quality Journal (REE&PQJ)
ISSN:3020-531X
Volume 26, July 2026

Short-Term Streamflow Forecasting for Run-Of-River Hydropower Plants using PCA-MLP and Local Linear Adjustment

J.D. Bautista-Izquierdo and L.A. Fernández-Jiménez

Department of Electrical Engineering
E.T.S.I.I., University of La Rioja
C/ San José de Calasanz 31– 26004 Logroño (Spain). Phone number: +0034 941 299473. E-mail: jebautis@unirioja.es

Abstract. The participation of renewable energy sources in power systems requires reliable short-term generation forecasts to ensure the efficient operation of the electricity market. Small hydropower plants (SHPPs), which represent a significant share of renewable generation, are particularly affected by hydrological and meteorological variability. This impact is especially pronounced in basins with moderate streamflow, seasonal intermittency, and lack of upstream regulation, as occurs in several unregulated sub-basins of the Ebro River. This study addresses the challenge of forecasting average daily flow in SHPPs operating under data-scarce conditions. A hybrid forecasting framework is proposed, combining physically based hydrological information and meteorological predictions within a machine learning model. Dimensionality reduction is applied to improve the robustness and stability of the model. In addition, a local linear post-processing adjustment is introduced, incorporating recent flow observations up to the day prior to the forecast, including intraday measurements from the same day. The proposed approach is computationally efficient and requires a low volume of data, showing a significant improvement in forecasting accuracy compared to standard methods. Its design makes it suitable for real-world applications as a preliminary step for forecasting electricity production and its use in electricity markets, where hydrological uncertainty is a common constraint.

Keywords. Small hydropower plants, Streamflow forecasting, Machine learning, PCA, Hydrological uncertainty.

1. Introduction

The transition toward decarbonized power systems has driven a sustained increase in the participation of renewable energy sources in the generation mix. This structural shift, while essential for mitigating climate change, introduces new operational challenges arising from the inherent variability and uncertainty of these resources. In this context, the availability of reliable short-term forecasting tools has become a key element to ensure the efficient operation of electricity markets, optimize generation scheduling, and maintain supply security.

Hydropower continues to play a fundamental role among renewable technologies, both due to its high efficiency and its rapid response capability to fluctuations in demand and other uncontrollable sources. Within this sector, small hydropower plants (SHPPs) represent a significant portion of distributed generation and contribute substantially to local and regional energy supply. However, unlike large facilities with regulating reservoirs, SHPPs typically operate under run-of-river conditions or with very limited storage capacity, making them highly dependent on the immediate availability of streamflow.

This direct dependence on water resources makes SHPPs especially sensitive to hydrological and meteorological variability. Factors such as the temporal distribution of precipitation, temperature, evapotranspiration, and basin characteristics strongly influence the daily available flow. This variability is particularly pronounced in basins with moderate flows, marked seasonal intermittency, and no upstream regulation, where hydrological response can be rapid and highly nonlinear. Such conditions are common in many unregulated sub-basins, such as those of the Ebro River, as the Oja River basin (Fig. 1), where high-quality historical data are often scarce.

Traditionally, streamflow prediction has been addressed using physically based or conceptual hydrological models. These models represent dominant hydrological processes and have been widely used in water resources planning and management. However, their practical application may be limited under data-scarce conditions, high uncertainty in meteorological inputs, or difficulties in parameter calibration, especially in small or poorly monitored basins.

In recent decades, machine learning techniques have opened new opportunities for hydrological forecasting. Many studies have explored artificial neural networks, support vector machines, tree-based methods, and deep learning models for streamflow estimation and prediction [1]. These approaches are excellent for capturing complex nonlinear relationships between inputs and outputs, often outperforming traditional physically based models.

Current research shows growing interest in hybrid models combining hydrological and meteorological information with advanced machine learning techniques. Variable selection, dimensionality reduction, and hyperparameter optimization strategies have been proposed to improve model robustness and generalization [2]. However, most studies focus on well-monitored basins or larger hydropower facilities with extensive historical records. In contrast, daily streamflow prediction for SHPPs in data-scarce areas remains a relevant and underexplored challenge.

Another trend in recent literature is the use of increasingly complex models, such as LSTM networks, sequence-to-sequence architectures, or transformer-based models, which have shown promising results in hydrological applications [3]. However, these approaches often require large datasets for training, high computational resources, and provide lower interpretability, limiting their operational implementation [4]. For SHPPs with limited technical resources and data availability, these constraints are particularly significant.
Therefore, it is necessary to develop forecasting methodologies that balance accuracy, robustness, and simplicity, and can operate reliably under data scarcity and hydrological uncertainty. Moreover, these approaches should integrate information from multiple meteorological, hydrological, and geographical sources, leveraging both physical system knowledge and data-driven flexibility, without excessively increasing computational complexity.
This work addresses this challenge by proposing a hybrid framework for daily streamflow forecasting in SHPPs, combining physically based hydrological information and meteorological predictors within a machine learning model. To improve model stability and generalization under high dimensionality and heterogeneous data, dimensionality reduction via principal component analysis (PCA) is applied. In addition, a local linear post-processing adjustment incorporates recent flow observations up to the day prior to the forecast, including intraday measurements, which is especially relevant from an operational perspective.
The proposed approach features a simple structure, low computational cost, and reduced reliance on historical data, showing significant improvement in streamflow prediction accuracy compared to standard methods. These characteristics make the developed framework particularly suitable for real-world applications in electricity markets with high renewable penetration, where hydrological uncertainty and limited data availability are common constraints.

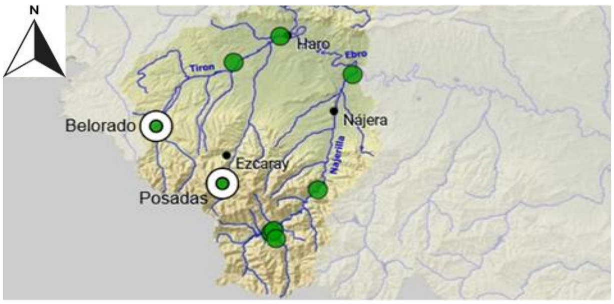

Fig. 1. Location of the study area, Oja River basin (CHEBRO).

2. Methodology

2.1. Data collection

Since the required data are not centralized in a single provider, it is necessary to obtain them from various agencies and institutes, at both national and international levels, each specializing in a specific type of information. This heterogeneity in data sources requires a preliminary process of collection, validation, and standardization, which is a key step for the proper development and training of the forecasting model.
The collected data constitute the fundamental input for the design and calibration of the proposed model. The consulted sources and organizations are listed below:

- MeteoGalicia – Xunta de Galicia
- NCAR – National Center for Atmospheric Research
- WRF – Weather Research and Forecasting Model
- CHE – Ebro River Basin Authority
- SAIH – Automatic Hydrological Information System
- IGN – National Geographic Institute
- CNIG – National Geographic Information Center
- EEA – Copernicus Land Monitoring Service
- SIOSE – Land Use Information System
- IGME – Geological and Mining Institute of Spain
- SIAR La Rioja – Agroclimatic Information System
- SNCZI IPE – Inventory of Dams and Reservoirs (Spain)

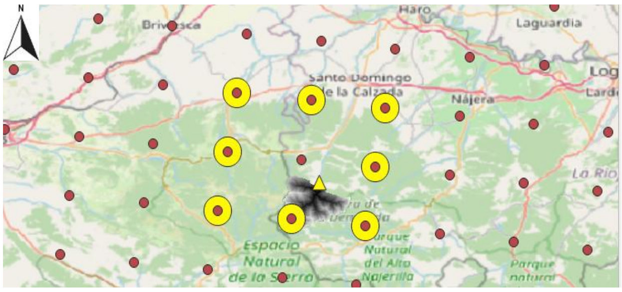

Fig.2. Model grid over the study basin (MeteoGalicia).

2.2. MeteoGalicia data acquisition

The meteorological variables used in this study come from the deterministic WRF-HIST model of MeteoGalicia, corresponding to the d02 domain, which provides hourly predictions on a regular grid over Galicia and adjacent regions. The data are distributed in NetCDF4 format through the MeteoGalicia THREDDS server.
Fig. 2 shows the d02 domain, which has a spatial resolution of approximately 12 km, making it suitable for basin-scale hydrological studies. It provides forecasted values for the key atmospheric variables shown in Table I, at regularly gridded geographic points at the aforementioned resolution (red points in the figure).
To automatically obtain the meteorological data, a Python script (Script 1) was developed to enable the systematic download of all daily files from the WRF-HIST model for a user-defined time interval. For each day, the URL corresponding to the file is dynamically generated and stored locally ensuring a continuous database:
"wrf_arw_det_history_d02_YYYYMMDD_0000.nc4".

2.3. Definition of the spatial prediction scheme

For each gauging station, a geographic reference point (latitude and longitude) was defined. From this point, the nearest node in the WRF grid was identified by minimizing the Euclidean distance in geographic coordinates.
Based on this node, three schemes are defined to capture WRF model prediction data:

P1. The prediction data are extracted from a single grid point (the nearest node).

P2. The prediction data are computed as the arithmetic mean of the values predicted at the four adjacent grid points (north, south, east, and west).

P3. Similar to P2 but using eight nodes surrounding the central node (points highlighted in yellow in Fig. 2).

This approach allows the evaluation of the effect of spatial smoothing on the representativeness of meteorological variables. The selected points were exported in Shapefile format (EPSG:4326), enabling visualization and validation in the geographical information system QGIS, which facilitates spatial inspection, overlay with the study basin, and verification of the prediction point locations.

2.4. *Extraction of hourly meteorological variables*

Using a second Python script (Script 2), the daily NetCDF files were processed, and forecasted hourly values of the following variables were extracted for each grid point:

Table I. MeteoGalicia netCDF hourly explanatory variables.

Variable	Unit	netCDF
Rainfall	(Kg/m^2)	prec
Convective precipitation	(Kg/m^2)	conv_prec
Snow precipitation	(Kg/m^2)	snow_prec
Shortwave radiation	(Wh/m^2)	swflx
Sea-level pressure	(Pa)	mslp
Air temperature	(K)	temp
Relative humidity	(%)	rh

The results were stored in a structured Excel file, where each row corresponds to one hour and each column corresponds to a variable at a specific WRF grid point. In parallel, a shapefile is generated containing the geographic data for each point which can be loaded into QGIS, allowing spatial inspection and verification that the selected nodes are correctly located within the study basin.

2.5. *Spatial hourly averaging (P3 Prediction)*

Since the P3 scheme (8 points) provides greater stability and reduces the spatial noise of the model, it was used as the selected meteorological scheme. For each variable, the spatial average was calculated, as defined in Equation (1), obtaining a single time series representative of the area surrounding the gauging station:

$$\mathrm{P3\_media\_\{var\}} = \frac{1}{8}\sum_{i=1}^{8} \mathrm{P3\_n\{i\}\_\{var\}} \qquad (1)$$

where P3_n{i}_{var} represents each variable (var with netCDF names extracted from Table I) from the MeteoGalicia netCDF file, and *i* corresponds to the index of the considered point.

2.6. *Temporal hourly-daily normalization*

The hourly time series were reindexed onto a continuous time axis, ensuring the explicit presence of all hours within the study period. This step is essential for the correct calculation of accumulated values, moving averages, and daily aggregations.

2.7. *Calculation of daily evapotranspiration*

Equation (2) shows how daily evapotranspiration, ET (mm/day or $kg/m^2/day$), was calculated using the Hargreaves–Samani method [5], based on daily mean, maximum, and minimum air temperatures (in °C). The hourly variable P3_media_swflx (Wh/m^2) was converted into daily solar radiation, R_a ($MJ/m^2/day$).

$$\mathrm{ET} = 0.023 \cdot \mathrm{R_a} \cdot (\mathrm{T_{mean}} + 17.8) \cdot \sqrt{\mathrm{T_{max-}T_{min}}} \qquad (2)$$

Considering the accumulated daily precipitation and subtracting the evapotranspiration ET, the effective rainfall P_ET_MetGal (Kg/m^2) is defined in Equation (3):

$$\mathrm{P\_ET\_MetGal} = \mathrm{P} - \mathrm{ET} \qquad (3)$$

where P (kg/m^2) represents daily rainfall, obtained from the accumulated hourly variable P3_media_prec [6].

Table II. Preprocessed daily explanatory variables.

Variable		Unit	DT	Interval	Measure
Year		year	O	[d]	--
Day_sine		--	O	[d]	--
Day_cosine		--	O	[d]	--
Q_SCS		m^3/s	F	[d-1,d+1]	M
Q_CHE		m^3/s	O	[d-15,d-1]	M
prec _MetGal		Kg/m^2	F	[d-30, d+1]	A
prec _CHE		L/m^2	O	[d-30, d-1]	A
snow_MetGal		Kg/m^2	F	[d-30,d+1]	A
temp_MetGal		K	F	[d-2,d+1]	M
rh_MetGal		%	F	[d-2,d+1]	M
swflx_MetGal		Wh/m^2	F	[d-1,d+1]	A
mslp_MetGal		Pa	F	[d,d+1]	M
P_ET_MetGal		Kg/m^2	F	[d-1,d+1]	A
CN_SCS		--	F	[d,d+1]	--
S_SCS		mm	F	[d,d+1]	A
Net_prec_SCS		mm	F	[d,d+1]	A
Q_12:00_d		m^3/s	O	[d]	H,M
Q[d+1]	Target	m^3/s	F	[d+1]	M
DT: data type, O: observed, F: forecasted, M: mean, A: accumulated and H: hourly					

2.8. *CHEBRO data acquisition*

Hydrological data from the Automatic Hydrological Information System of the Ebro River Basin (SAIH Ebro), operated by the Ebro Hydrographic Confederation (CHEBRO), were used as reference for observed river flows and precipitation across the study basin. Historical data were extracted in XLSX format, including daily mean, maximum, and minimum river flows, as well as daily accumulated precipitation, with the hour of occurrence recorded for daily flow extrema. The measurements originate from the river gauging station within the basin. Using this information, a representative intraday flow at 12:00 h, denoted as Q_12:00_d (Table II), was reconstructed or interpolated. Although the intraday recorded value is not available, this interpolated variable provides valuable intraday information immediately prior to predictions. In real-time operation, the model could incorporate the observed flow of day d directly. These data were employed as observed variables for model calibration and validation, as well as for computing effective rainfall and generating daily time series of hydrological and meteorological variables.

2.9. *Curve Number estimation (CN)*

For the characterization of the watershed, raster data from IGN, IGME, EDM, and NDVI are used (Fig. 3), processed through a Python script integrated with QGIS. These data, corresponding to digital elevation models (DEM), enable the extraction of the main physical and morphometric parameters of the watershed, such as land use, soil type, slope, and vegetation cover. One of the most relevant parameters is the Curve Number (CN) from the SCS method (see next section), calculated under different soil moisture conditions: dry (CN-I), intermediate (CN-II), and very wet (CN-III). From these CN values, the maximum soil retention, S, is also derived, which allows estimation of net precipitation (P_n) and, through a physical model, the generation of streamflows that simulate the real behavior of the watershed in response to precipitation events. The main purpose of the CN in this study is to estimate the expected mean flow on future days, Q_SCS (Table II), which enables integration of watershed and precipitation information into hydrological prediction. This comprehensive characterization facilitates the construction of thematic maps and parameter series that are incorporated into the hydrological model to compute runoff, effective rainfall, and daily-scale flow simulations.

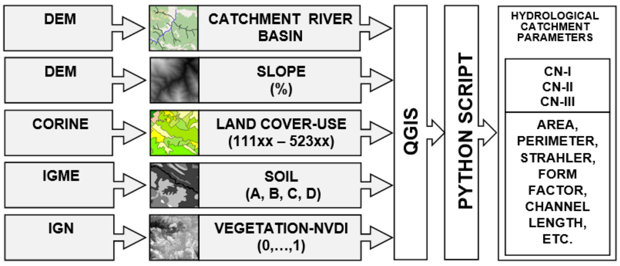

Fig. 3. Schematic workflow for basin characterization using raster data and processing in QGIS with Python (Script 3).

2.10. *Daily mean estimated discharge Q_SCS*

This section presents the estimation of daily mean discharge using the Soil Conservation Service (SCS) method, also known as the Curve Number (CN) method [5]. This procedure enables the calculation of watershed surface runoff from recorded precipitation while also considering soil moisture conditions and land use and land cover characteristics. To do this, the Antecedent Moisture Condition (AMC) is first determined, which is classified into three categories: dry (AMC I), normal (AMC II), and wet (AMC III). These conditions are established based on the accumulated precipitation over the previous five days and according to the time of year (growing period or dormant period), as shown in Table III. This classification is essential because it directly influences the Curve Number value used in the calculations.

Table III. Antecedent Moisture Conditions (AMC).

Plant period	AMC I (Dry)	AMC II (Normal)	AMC III (Wet)
Dormant	< 13 mm	13–32 mm	> 32 mm
Growing	< 35 mm	35–52 mm	> 52 mm
Previous 5-day accumulated precipitation (mm)			

Once the moisture condition has been defined, the Curve Number (CN) corresponding to the watershed is adjusted. The base CN value, corresponding to normal conditions (CN-II), which is one of the first results calculated by Script 3, is corrected using Equation (4) to obtain the equivalent values for dry or wet conditions (according to AMC). With the adjusted CN, the parameter S is calculated, which represents the maximum potential soil retention capacity, through Equation (5):

$$\mathrm{CN}_{\mathrm{I-II-III}} = \begin{cases} \text{Dry} \rightarrow & \dfrac{4.2 \cdot \mathrm{CN-II}}{10 - 0.058 \cdot \mathrm{CN-II}} \\ \text{Normal} \rightarrow & \mathrm{CN-II} \\ \text{Wet} \rightarrow & \dfrac{23 \cdot \mathrm{CN-II}}{10 - 0.058 \cdot \mathrm{CN-II}} \end{cases} \quad (4)$$

$$S = \frac{25400}{\mathrm{CN}_{\mathrm{I-II-III}}} - 254 \quad (5)$$

Subsequently, the effective precipitation or daily direct runoff (P_{net}) is determined with (6). This calculation considers the daily accumulated precipitation (P), the daily evapotranspiration (ET), and the previously obtained parameter S. If the available precipitation is less than 0.2S, it is assumed that no runoff is generated; otherwise, the SCS formula is applied to estimate the volume of water that contributes to surface runoff.

Finally, using the net precipitation calculated previously and the watershed area estimated through a Python script (Script 3) developed within the QGIS environment, the total runoff volume generated is calculated. This runoff is then assumed to be uniformly distributed throughout the day, and the daily mean estimated discharge value, referred to as Q_SCS, is obtained. In this way, it is possible to evaluate the hydrological behavior of the watershed on a daily scale, that is, the effective precipitation based on meteorological data and physical terrain parameters (6),

$$P_{net} \begin{cases} 0 & (P - ET) < 0.2 \cdot S \\ \dfrac{((P - ET) - 0.2 \cdot S)^2}{(P - ET) - 0.8 \cdot S} & (P - ET) \geq 0.2 \cdot S \end{cases} \quad (6)$$

where P is the daily accumulated precipitation (mm), ET is daily evotranspiration (mm), S is the soil's maximum potential retention (mm), and P_{net} is the effective precipitation or resulting direct runoff (mm).

The daily mean discharge estimated by the SCS method is obtained from the previously calculated effective precipitation and the watershed area:

$$Q\_SCS = \frac{P_{net} \cdot A_{basin} \cdot 1000}{24 * 3600} \quad (7)$$

where Q_SCS represents the estimated daily mean discharge in m^3/s, A_{basin} is the watershed area in km^2, and 24·3600 is the total number of seconds in a day. Using expression (7), the total runoff volume generated during the day is converted into a mean discharge value.

2.11. *Other explanatory variables*

Sine and cosine transformations of the date (Day_sine, Day_cosine) are used to capture annual seasonality, with d taking values from 1 to 365.

$$\begin{aligned} \text{Day\_sine} &= \sin\left(\frac{d \cdot 2\pi}{365}\right) \\ \text{Day\_cosine} &= \cos\left(\frac{d \cdot 2\pi}{365}\right) \end{aligned} \tag{8}$$

Additionally, the variables associated with meteorological forecasts or observations (prec_MetGal, prec_CHE, snow_MetGal, temp_MetGal, rh_MetGal, swflx_MetGal, and mslp_MetGal), listed in Table II, were used as explanatory variables, together with their accumulated values at 1, 2, 3, 4, 5, 15, and 30 days. In this way, the total number of explanatory variables used in this study was 56.

3. Forecasting model

In the following paragraphs, the model that achieved the best forecasting results is presented. Different model families were evaluated, ranging from fully analytical models to black-box approaches such as neural network-based models. After multiple tests, the results showed that combining analytical and neural approaches provided the best predictive performance. This hybrid model is the one described in the following sections. For the sake of operational simplicity, a single model was adopted for all weather conditions.

3.1. Model structure

A multilayer perceptron (MLP) is used as the predictive core. Prior to training, the input data are transformed using principal component analysis (PCA) to reduce dimensionality and improve the stability and generalization of the model in the presence of highly heterogeneous variables [7]. The architecture consists of two hidden layers, with the exact number of neurons determined through optimization using a genetic algorithm (GA) and cross-validation (Fig. 4), ensuring that physical constraints guide the input space while the MLP handles nonlinear mapping.

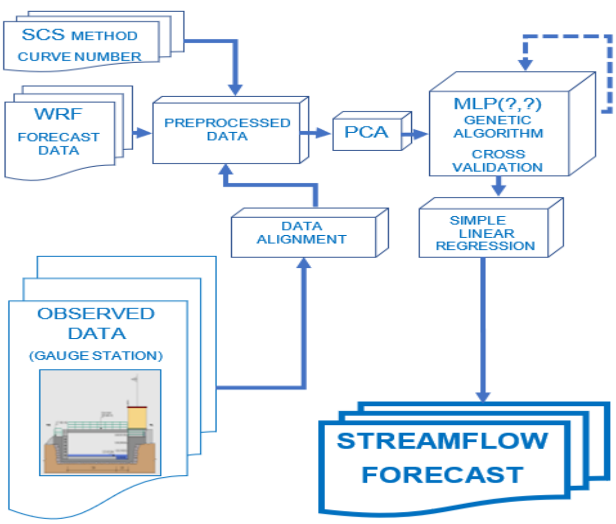

Fig. 4. Model workflow.

3.2. Principal Components Analysis (PCA)

Before training the neural network, the 56 preprocessed explanatory variables are standardized by removing the mean and scaling to unit variance to ensure that every feature contributes on a comparable scale. After this, PCA is applied with scikit-learn to reduce dimensionality and remove redundancy among highly correlated variables. Only the first principal components that together explain at least 60% of the total variance are retained (Table IV), ensuring that most of the relevant information is preserved. Table IV also presents, on the right-hand side, the original explanatory variables with the highest weights in the first principal component (PC1).

This dimensionality-reduction step produces a compact and decorrelated feature set. Importantly, the MLP is not fed with the original variables but with the selected principal components (PCs), which serve as its actual input features. Working directly with the PCs makes the model more efficient, less prone to overfitting, and more robust against multicollinearity present in the original dataset.

Table IV. PCA: explained variance and variable loadings.

PC	EV (%)	Cumulative variance (%)	Variable	Loading
PC1	32.47	32.47	**P_ET_MetGal-1d**	**0.0994**
PC2	11.34	43.81	rain_CHE-2d_ac	0.0987
PC3	9.86	53.67	rain_CHE-5d_ac	0.0985
PC4	7.08	**60.74**	rain _CHE-30d_ac	0.0976
PC5	6.04	66.79	snow_MetGal-1d	0.0945
PC6	4.45	71.24	swflx_MetGal	0.0939
PC7	3.06	74.30	rain _CHE-1d	0.0933
PC8	2.67	76.97	rain _MetGal-15d_ac	0.0930
PC9	2.33	79.30	snow_MetGal-5d_ac	0.0925
PC10	2.03	81.33	rain _CHE-3d_ac	0.0919

3.3. Multi Layer Perceptron (MLP)

The core model is a two-hidden-layer MLP neural network (MLP Regressor). The hidden layers use ReLU activation to capture nonlinear relationships between the PCs, while the output layer is linear, preserving the continuity of the predicted flow and truncating negative values to zero.

The architecture is optimized using a GA, evaluating different configurations with 5-fold cross-validation and using the Nash–Sutcliffe Efficiency (NSE) as the performance metric.

3.4. Local linear adjustment (Post-Processing)

To correct pointwise deviations of the MLP and bring predictions closer to observed values, a local linear adjustment is applied based on the most recent observed data.

- An observation window (WINDOW) is defined, containing the most recent observed discharge values and the midday discharge of the current day.
- A local linear regression is constructed between the MLP-predicted values and the observed values within the window.
- The final prediction is adjusted by interpolating along the regression line, ensuring temporal continuity and eliminating potential negative values.

This approach allows incorporating intraday information available up to 12 hours before the start of day d+1, increasing accuracy without compromising model simplicity.

3.5. Training and validation

- Training period: all days in 2012–2021
- Test period: all days in 2022–2025
- Training/validation split: 80–20% with K-fold cross-validation (K=5) to tune MLP hyperparameters.

Table V shows error indicators of the predictions obtained with the hybrid model (PCA-MLP+Local Linear Adjustment) compared with those from different models.

4. Results

The results obtained with the proposed model are compared with those from a set of benchmark forecasting approaches, including a naïve model, several time-series models, and linear regression models using the same explanatory variables. The naïve model uses the most recent observed value as the forecast. The results show that purely statistical models, such as the ARIMA and SARIMA variants, achieve reasonable performance, with coefficients of determination (R^2) around 0.85 and RMSE errors close to 1.0 m^3/s, without requiring regime-specific calibration.

Table. V. Results obtained with different techniques

Technique	R^2	RMSE (m^3/s)	MAE (m^3/s)
Naïve	0.833	1.081	0.287
Simple Linear Regression	0.840	1.058	0.323
Multiple Linear Regression	0.895	0.806	0.337
ARIMA(1,0,0)	0.840	1.058	0.323
ARIMA(2,0,0)	0.841	1.055	0.323
ARIMA(0,0,1)	0.615	1.639	0.980
ARIMA(0,0,2)	0.736	1.357	0.722
ARIMA(1,0,1)	0.841	1.053	0.322
$SARIMA_{(2,1,2)x(1,0,1,15)}$	0.846	1.353	0.306
PCA_MLP_LINEAL(27,24)	**0.957**	**0.358**	**0.135**

However, these models exhibit clear limitations in capturing nonlinear relationships between meteorological variables and the hydrological response of the basin. Similarly, linear regression models, although they slightly improve performance compared to a naïve predictor, show limited ability to generalize in complex hydrological scenarios.

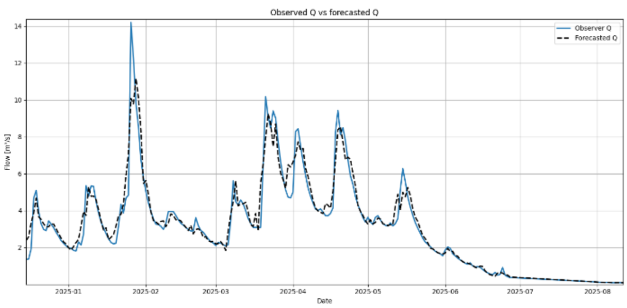

Fig. 5. Forecasted vs observed daily average streamflow.

The best overall performance is obtained with the proposed PCA_MLP_LINEAL model, which combines a neural network to capture nonlinearities with a linear term to preserve physical consistency and stability. The optimal values obtained through GA optimization and cross-validation were 27 neurons in the first hidden layer and 24 in the second. This model achieves an R^2 of 0.95, an RMSE of 0.35 m^3/s, and an MAE of 0.13 m^3/s, substantially outperforming both purely statistical models. Its reduced RMSE and MAE values indicate a significant improvement in prediction robustness, especially under variable hydrological conditions. Fig. 5 plots the predicted mean streamflow values obtained with the proposed model and the real values during eight months in the test period.

5. Conclusion

This study presents an integrated framework for the prediction of daily streamflow Q[d+1], combining high-resolution meteorological variables from the MeteoGalicia WRF model, a physically consistent hydrological preprocessing based on the SCS–CN method, and a broad set of statistical and machine learning models. The full automation of the data workflow through Python scripting has enabled the development of a reproducible, scalable system free from manual bias, capable of coherently integrating meteorological, spatial, and temporal information.

The results demonstrate that the integration of meteorological variables, physically derived hydrological variables (effective precipitation, antecedent moisture, CN), and hybrid learning techniques constitutes a highly effective strategy for short-term streamflow prediction.

Overall, the study shows that the real improvement does not lie solely in the use of advanced models, but rather in the construction of a physically informed, spatially consistent, and fully automated data pipeline, capable of transforming meteorological forecasts into reliable and actionable hydrological information.

6. References

[1] Z. M. Yaseen, A. El-Shafie, O. Jaafar, H. A. Afan and K. N. Sayl, “A review of state-of-the-art artificial intelligence applications in hydrology,” Water Resources Management (2015). Vol. 29, pp. 4281–4305.

[2] O. Kisi and J. Shiri, “Improving accuracy of river flow forecasting using LSSVR with gravitational search algorithm,” Water Resources Management (2014). Vol. 28, pp. 2073–2090.

[3] J. Zhang, H. Chen and Y. Li, “Streamflow prediction using an integrated methodology based on convolutional neural network and long short-term memory networks,” Journal of Hydrology (2018). Vol. 558, pp. 1–12.

[4] Q. Duan et al., “A state-of-the-art review of long short-term memory models with applications in hydrology and water resources,” Hydrology and Earth System Sciences (2019). Vol. 23, pp. 1–25.

[5] V. Mockus, “Estimation of direct runoff from storm rainfall,” in National Engineering Handbook, Section 4: Hydrology, U.S. Department of Agriculture, Soil Conservation Service, Washington DC, USA (1964).

[6] Natural Resources Conservation Service (NRCS), National Engineering Handbook, Part 630: Hydrology, U.S. Department of Agriculture, Washington DC, USA (2004).

[7] S. Ahmad et al., “Short-term hydropower generation prediction model for run-of-river hydropower plants considering hydrometeorological factors,” Renewable Energy (2021). Vol. 170, pp. 613–627.