



OPEN Short-term forecasts of streamflow in the UK based on a novel hybrid artificial intelligence algorithm

Fabio Di Nunno, Giovanni de Marinis & Francesco Granata✉

In recent years, the growing impact of climate change on surface water bodies has made the analysis and forecasting of streamflow rates essential for proper planning and management of water resources. This study proposes a novel ensemble (or hybrid) model, based on the combination of a Deep Learning algorithm, the Nonlinear AutoRegressive network with exogenous inputs, and two Machine Learning algorithms, Multilayer Perceptron and Random Forest, for the short-term streamflow forecasting, considering precipitation as the only exogenous input and a forecast horizon up to 7 days. A large regional study was performed, considering 18 watercourses throughout the United Kingdom, characterized by different catchment areas and flow regimes. In particular, the predictions obtained with the ensemble Machine Learning-Deep Learning model were compared with the ones achieved with simpler models based on an ensemble of both Machine Learning algorithms and on the only Deep Learning algorithm. The hybrid Machine Learning-Deep Learning model outperformed the simpler models, with values of R^2 above 0.9 for several watercourses, with the greatest discrepancies for small basins, where high and non-uniform rainfall throughout the year makes the streamflow rate forecasting a challenging task. Furthermore, the hybrid Machine Learning-Deep Learning model has been shown to be less affected by reductions in performance as the forecasting horizon increases compared to the simpler models, leading to reliable predictions even for 7-day forecasts.

Abbreviations

1	Indicator function
1st Q	First quartile of the daily streamflow rate
2nd Q	Second quartile of the daily streamflow rate
3rd Q	Third quartile of the daily streamflow rate
4th Q	Fourth quartile of the daily streamflow rate
ANFIS	Adaptive Neuro Fuzzy Inference System
ANN	Artificial Neural Network
AI	Artificial intelligence
b	Bias in NARX model
BO	Bayesian optimization
BR	Bayesian regularization
BI-LSTM	Bidirectional Long Short-Term Memory
BRF	Boruta Feature Selection
BPNN	Back-Propagation Neural Network
BRNN	Bayesian Regularization Neural Network
BWNN	Bootstrap Wavelet Neural Network
CNN	Convolutional Neural Network
CV_Q	Coefficient of variation of the daily streamflow rate
DL	Deep Learning
ESP	Ensemble Streamflow Prediction
f_i	Sigmoid activation function
FFNN	Feed-Forward Neural Network
GRU	Gated Recurrent Unit
GWO	Grey Wolf Optimization
h	Hidden nodes in NARX model

Department of Civil and Mechanical Engineering (DICEM), University of Cassino and Southern Lazio, Via Di Biasio, 43, 03043 Frosinone, Cassino, Italy. ✉email: f.granata@unicas.it

IWD	Intelligent Water Drop
LASSO	Least Absolute Shrinkage and Selection Operator
LM	Levenberg–Marquardt
LSTM	Long Short-Term Memory
ML	Machine Learning
MLR	Multiple-Linear Regression
n	Number of neurons in NARX model
NARX	Nonlinear AutoRegressive network with eXogenous inputs
OLS	Ordinary Least Squares
PSO	Particle Swarm Optimization
$\overline{Q_A}$	Mean streamflow rate
Q_A^i	Measured streamflow rate for the i th data
Q_P^i	Predicted streamflow rate for the i th data
$Q_{\max}$	Maximum daily streamflow rate
Q_{mean}	Mean daily streamflow rate
Q_{median}	Median daily streamflow rate
$Q_{\min}$	Minimum daily streamflow rate
$R(t_{RF})$	Impurity at each node in RF model
RF	Random Forest
SCG	Scaled Conjugate Gradient
$\text{sgn}(\cdot)$	Sign Function
Skew_Q	Skewness of the daily streamflow rate
SVM	Support Vector Machine
SVM-LF	Support Vector Machine with Linear kernel function
SVM-RF	Support Vector Machine with Radial basis kernel function
SVR	Support Vector Regression
s	Number of samples
t	Forecast horizon
t_{RF}	Node in the RF model
UK	United Kingdom
w	Weight in NARX model
WANN	Wavelet Artificial Neural Network
WSVM	Wavelet Support Vector Machine
$x(t)$	Value of the exogenous input at time t in the NARX model
y_i	Target variable in the i th unit in the RF model
y_m	Mean target variable in the node t_{RF}
$y(t)$	Target at time t in the NARX model
σ_Q	Standard deviation of the daily streamflow rate

River discharge forecasting plays an essential role in flood protection and water resources planning and management. River flows are increasingly influenced by the climate changes observed in recent decades, which are leading to increasingly frequent flood and drought events¹. In this scenario, optimal water resource management cannot disregard the prediction of river flows in the short and long term. However, while for the long term the considerable uncertainty of forecasts means that only trends can be reliably defined, for the short term it is possible to obtain even very accurate forecasts. These predictions can be conducted using different approaches, including physically based models, which consist of various mathematical equations used to describe hydrological processes^{2,3}, and conceptual models, which describe the same processes based on simplified equations and empirical relationships between parameters⁴. However, the high uncertainty and complexity associated with hydrological processes and weather-climate factors affecting river basins have led researchers to increasingly use data-driven approaches, in particular Artificial Intelligence (AI) algorithms, which guarantee fast processing without the need to define complex analytical relationships between input and target variables⁵. AI algorithms have been widely applied in recent years to tackle various hydrological problems^{6,7}. Among these, several Machine Learning (ML) algorithms were used for the prediction of streamflow rate^{8–12}. In addition, to improve streamflow predictions, in the last few years researchers have moved towards the development of so-called hybrid or ensemble models, based on the combination of different individual ML and optimization algorithms. Li et al.¹³ compared three different ML algorithms: Back-Propagation Neural Network (BPNN), Support Vector Regression (SVR), and Adaptive Neuro Fuzzy Inference System (ANFIS), for the daily streamflow rate prediction for the Yuetan Basin, China. In particular, the authors applied the wavelet threshold de-noising method as pre-processing for time series. Then, both BPNN and SVR were combined with the Particle Swarm Optimization (PSO) algorithms. They showed how the PSO-SVR model showed a better overall performance compared to both PSO-BPNN and ANFIS models. Pham et al.⁷ proposed a hybrid model based on a ML algorithm, the Multi-Layer Perceptron (MLP), and an Intelligent Water Drop optimization algorithm (MLP-IWD) for the river flow rate forecasting of the Vu Gia Thu Bon River, Vietnam. The authors compared the predictions made with the individual MLP algorithm and the ensemble MLP-IWD, showing how hybridization led to a marked increase in performance. Saraiva et al.¹⁴ presented a comparative analysis of two ML models: Artificial Neural Network (ANN) and Support Vector Machine (SVM), coupled with wavelet transform and data resampling with the bootstrap method, applied for the daily streamflow rate forecasting for Sobradinho Reservoir, Brazil. The authors showed that the best combination was the BWNN, obtained combining Bootstrap (B), Wavelet (W) and Neural Network (NN),

highlighting the advantages of the ensemble approach. Tyralis et al.¹⁵ developed a super ensemble model for one-step-ahead daily streamflow forecasting on 511 basins located in USA, based on 10 different ML algorithms. The super ensemble learning algorithm outperformed all individual ML algorithms, with, however, NN which provided the best prediction among the 10 individual algorithms. Kumar et al.¹⁶ compared the performance of two data-driven techniques, a Wavelet ANN (WANN) and a SVM with linear and radial basis kernel functions (SVM-LF and SVM-RF), for the daily discharge prediction of a Perennial River, India. The authors showed how SVM-RF outperformed both WANN and SVM-LF models. Kumar et al.¹⁷ also compared the performance of five different data-driven techniques: ANN, WANN, SVM, Wavelet SVM (WSVM) and Multiple-Linear Regression (MLR), for the forecasting of daily suspended sediment concentration in Indian Rivers, with the WSVM that outperformed the other four techniques.

Moreover, as the great potential of Deep Learning (DL) algorithms in the prediction of time series is now well known, a number of researchers have developed streamflow prediction models based on them in recent years. Fu et al.¹⁸ proposed a DL model based on LSTM to predict the streamflow of the Kelantan River, Malaysia. They compared the performance of the LSTM model with that of a classical neural network with back-propagation and found a higher accuracy of the LSTM model in predicting both regular flow and rapid fluctuations in the dry and rainy seasons, respectively. Le et al.¹⁹ presented a comparative analysis of six DL models, including: Feed-Forward Neural Network (FFNN), Convolutional Neural Network (CNN), and four Long Short-Term Memory (LSTM) -based models, applied for streamflow forecasting in the Red River basin, Vietnam. They also compared the performance of two simpler LSTM and Gated Recurrent Unit (GRU) models, with only one hidden layer, with two more complex models, the Stacked-LSTM model and the Bidirectional LSTM (Bi-LSTM) ones. The authors indicated how the LSTM models outperformed both FFNN and CNN models. However, the higher complexity of the Stacked-LSTM and Bi-LSTM models did not lead to a significant performance increase compared to the simpler LSTM models. Ahmed et al.²⁰ proposed a hybrid model based on the LSTM algorithm, used in conjunction with the Boruta Feature Selection (BRF) algorithm for the optimal choice of predictors, and applied it to the prediction of streamflow forecasting in six rivers in the Murray Darling Basin, Australia. They compared the performance of the BRF-LSTM model with other ML/DL -based models: individual LSTM, GRU, Recurrent Neural Network (RNN) and SVR, with the BRF-LSTM model that outperformed all the other models. Granata et al.²¹ proposed a comparison between two different models for the daily streamflow forecasting: an ensemble model based on Random Forest (RF) and Multilayer Perceptron (MLP), hybridized using the Stacking ML technique, and a Bi-directional Long Short-Term Memory (Bi-LSTM) network, where for both the hyperparameters were optimized based on a Bayesian process. The authors showed how the ensemble model outperformed the Bi-LSTM network in predicting peaks of flow rates, with also computation times significantly shorter. Wegayehu and Muluneh²² also compared three DL algorithms: Stacked-LSTM, Bi-LSTM and GRU, with the classical MLP network for one-step daily streamflow forecasting for the rivers Abay and Awash, Ethiopia. They showed how both MLP and GRU algorithms outperform S-LSTM and Bi-LSTM on a nearly equal basis. A comprehensive review of the hybrid artificial intelligence and optimization modelling for streamflow forecasting was provided by Hassan Ibrahim et al.²³.

Current literature, including a recent study by the authors mentioned above²¹, shows that reliable streamflow prediction models can be obtained using both hybrid ML and DL algorithms. Hence the idea of a possible ML-DL hybridisation with the aim of improving forecasts for both periods with ordinary flow rates and during flood events. Moreover, a further essential aspect is the forecasting horizon, which is a key element in the management of flood events. Accordingly, the performances of the developed models were assessed for forecast horizons up to 7 days. In this work, a novel prediction model was therefore developed based on the hybridization of a particular DL-RNN algorithm, the Nonlinear AutoRegressive network with eXogenous inputs (NARX), with the two algorithms RF and MLP. To the authors' knowledge, no study in the literature proposes a hybrid model based on NARX, MLP and RF for the streamflow rate forecasting. NARX networks have proven to be a valuable tool for forecasting time series of several hydrological quantities²⁴. On the other hand, RF and MLP considered individually do not represent excellent solutions to the problem of forecasting hydrological time series, however, their combination can in some cases even outperform a very powerful algorithm such as LSTM networks²¹.

From this perspective, the prediction made with the hybrid NARX-MLP-RF model were compared with the ones achieved with both a model based on the single NARX algorithm and another based on the hybridization of MLP and RF. Model training, testing and subsequent comparisons were conducted as part of a large regional study, which considered the daily flow rates of 18 watercourses throughout the United Kingdom (UK). The regional scope of the comparative study represents a further innovative aspect, as UK is characterized by basins with both very different extents and characteristics of rainfall and flow regimes. Therefore, this study can provide insights into the usefulness of implementing more or less complex hybrid models depending on the features of each river.

Materials and methods

Case studies and dataset. The catchment areas of the 18 rivers investigated in this study cover a significant and varied portion of the UK territory, from Scotland, where the Dee, Deveron, Spey, Tay, Nith, Teviot and Tweed rivers were analyzed, to England, where the Thames, Test, Tamar, Trent Bure, Ribble and Leven rivers were considered, and finally to Wales, where the Dee, Severn, Teifi and Wye rivers were studied (Fig. 1).

For each measurement station, the daily cumulative precipitation and average river flow rate from January 1, 1961, to December 31, 2017, were considered. Catchment area of each basin and daily streamflow statistics were reported in Table 1. Figure 2 shows the average annual precipitation and the average annual discharge for each measuring station. The rivers investigated show considerable variability in terms of:

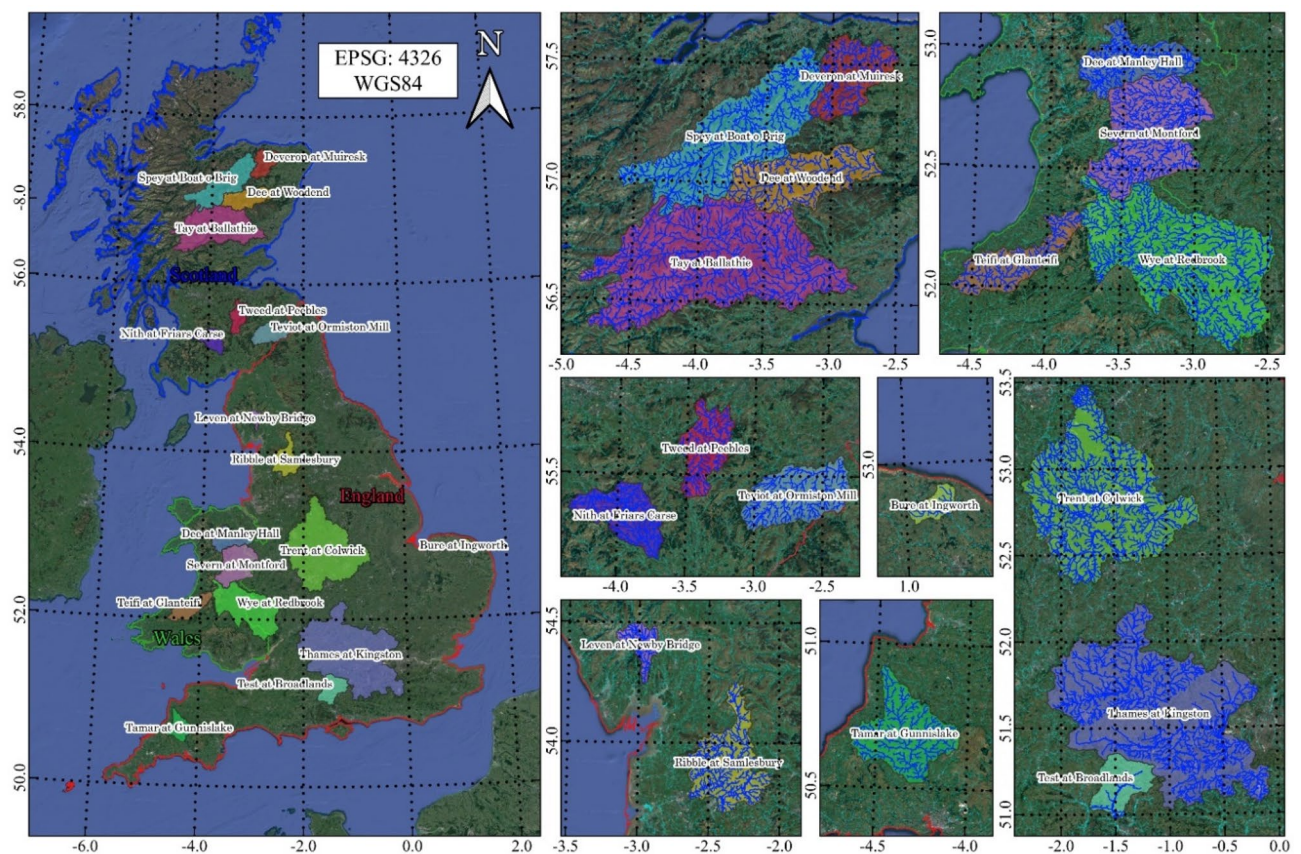


Figure 1. Location of the basins in UK. Maps created using the Free and Open Source QGIS²⁵.

	Bure at Ingworth	Dee at Manley Hall	Dee at Woodend	Deveron at Muirresk	Leven at Newby Bridge	Nith at Friars Carse	Ribble at Samesbury	Severn at Montford	Spey at Boat o Brig	Tamar at Gunnislake	Tay at Ballathie	Teifi at Glantefi	Test at Broadlands	Teviot at Ormiston Mill	Thames at Kingston	Trent at Colwick	Tweed at Peebles	Wye at Redbrook
Area (km ²)	161	1009	1381	962	248	798	1146	2028	2854	920	4589	898	1035	1122	9937	7472	699	4019
$Q_{\max}$ (m ³ /s)	1.15	31.08	37.76	16.88	14.30	28.50	34.12	43.66	65.71	22.45	175.25	28.82	10.92	20.77	62.36	84.71	16.05	73.13
Q_{median} (m ³ /s)	1.01	19.24	27.23	11.43	9.81	15.89	16.90	24.20	50.20	11.70	135.70	18.19	9.58	12.15	37.90	59.30	10.26	44.70
$Q_{\min}$ (m ³ /s)	11.8	521.0	860.8	387.7	191.0	467.1	765.0	462.0	1031.0	484.0	1965.0	447.2	36.8	554.8	581.0	982.0	306.7	781.0
Q_{min} (m ³ /s)	0.38	2.75	3.54	2.06	0.11	1.15	1.88	1.71	11.28	0.58	23.07	0.73	3.78	1.41	0.01	14.70	1.85	3.43
σ_Q (m ³ /s)	0.60	30.86	36.38	19.14	13.70	34.59	46.42	50.01	52.15	28.86	140.47	30.36	4.87	27.02	68.31	72.59	17.16	80.97
Skew _Q	0.72	1.15	0.87	0.85	0.98	1.09	1.11	1.17	0.89	1.12	0.84	1.05	0.82	0.96	1.07	1.05	1.01	1.05
CV _Q	0.52	0.99	0.96	1.13	0.96	1.21	1.36	1.15	0.79	1.29	0.80	1.05	0.45	1.30	1.10	0.86	1.07	1.11
1st Q	0.79	10.55	17.03	6.74	4.25	7.36	8.60	11.60	33.54	5.09	77.33	8.47	7.39	6.15	14.90	39.70	5.93	22.44
2nd Q	1.01	19.24	27.23	11.43	9.81	15.89	16.90	24.20	50.20	11.70	135.70	18.19	9.58	12.15	37.90	59.30	10.26	44.70
3rd Q	1.33	40.21	45.05	19.87	20.20	35.94	39.80	54.50	79.41	27.70	227.50	39.22	13.25	24.13	82.00	99.40	19.59	88.80
4th Q	11.8	521.0	860.8	387.7	191.0	467.1	765.0	462.0	1031.0	484.0	1965.0	447.2	36.8	554.8	581.0	982.0	306.7	781.0

Table 1. Catchment area and streamflow rate statistics for each basin. where: Area = catchment area, Q_{mean} = mean daily streamflow rate, Q_{median} = median daily streamflow rate, Q_{max} = maximum daily streamflow rate, Q_{min} = minimum daily streamflow rate, σ_Q = standard deviation of the daily streamflow rate, Skew_Q = skewness of the daily streamflow rate, CV_Q = coefficient of variation of the daily streamflow rate, 1st Q = first quartile of the daily streamflow rate, 2nd Q = second quartile of the daily streamflow rate, 3rd Q = third quartile of the daily streamflow rate, 4th Q = fourth quartile of the daily streamflow rate.

- Catchment area, ranging from 161 km², for Bure at Ingworth (eastern England), to 9937 km² for Thames at Kingston (southern England).
- Precipitation over the catchment area, ranging from an average annual precipitation (P_{annual}) of 696 mm, for Bure at Ingworth, to 2277 mm for Leven at Newby Bridge (northern England). Low P_{annual} values were also observed for Thames at Kingston and Trent at Colwick, in southern and central England, equal to 723 mm and 769 mm, respectively, while high P_{annual} values were observed for the Scotland rivers of Tay at Ballathie (northern Scotland) and Nith at Friars Carse (southern Scotland), with P_{annual} of 1499 mm and 1533 mm, respectively.

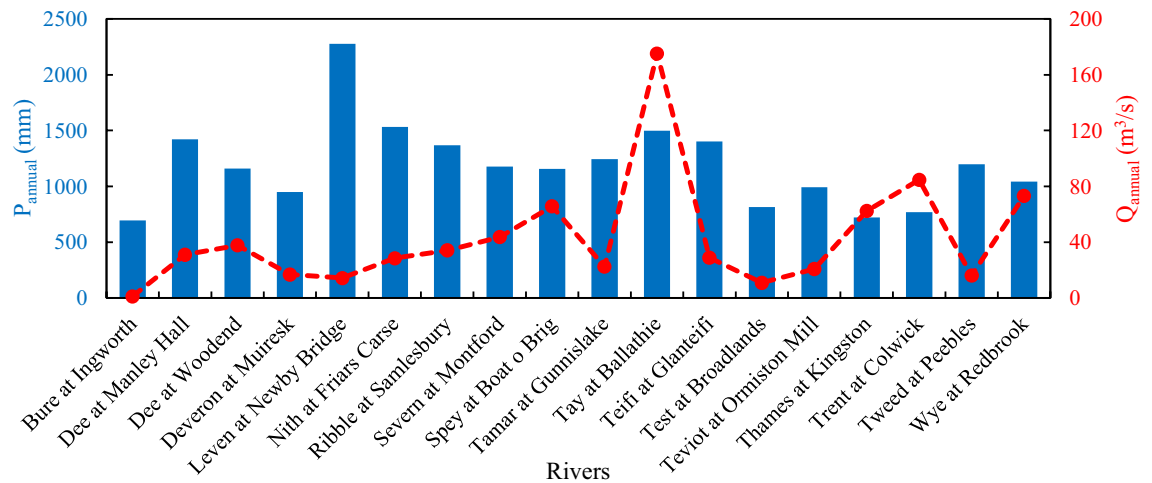


Figure 2. Average annual precipitation and streamflow rate.

- Streamflow rate: the lowest average annual discharge Q_{annual} was observed for Bure at Ingworth, equal to $1.15 \text{ m}^3/\text{s}$, while the highest Q_{annual} was observed for Tay at Ballathie, equal to $175.25 \text{ m}^3/\text{s}$. It should be noted that, despite Thames at Kingston has the largest catchment area of the 18 rivers, a Q_{annual} of $62.36 \text{ m}^3/\text{s}$ was observed, which was in line or even lower than other rivers with much smaller basins but with higher precipitations.

Forecasting algorithms. Three artificial intelligence algorithms, NARX, MLP and RF, were considered to develop models for predicting stream flows. Subsequently, the NARX-MLP-RF hybrid model was developed in order to obtain even more accurate predictions and was compared with both the MLP-RF hybrid model and the models based on the individual algorithms. The combination of algorithms was achieved by means of the stacking technique, which allows hybrid models to be developed from multiple regression or classification models²⁶. Specifically, individual models were first developed on the training dataset, then, based on the results of each model, a meta-learner was employed to develop the hybrid model. The Elastic Net algorithm²⁷ was chosen as the meta-learner in the present study. Elastic Net is a combination of two widely used regularized variants of linear regression: the Least Absolute Shrinkage and Selection Operator (LASSO) and the Ridge Regression. The main difference between LASSO and Ridge is represented by the penalty (or regularization) term. LASSO uses the L_1 regularization, with the aim of selecting the largest number of explanatory variables by introducing an absolute penalty to Ordinary Least Squares (OLS) regression. The L_1 regularization imposes sparsity among the coefficients making the fitted model more interpretable. Ridge uses the L_2 regularization, which also introduces a penalty in the OLS formulation, penalizing the square weights rather than the absolute ones. Moreover, the L_2 regularization limits the size of the coefficient vector. Elastic Net represents an optimal trade-off between Ridge and LASSO, with a penalty term which is a mix of the L_1 and L_2 regularizations²⁸, allowing to keep the feature selection quality from the LASSO penalty as well as the effectiveness of the Ridge penalty²⁷. The parameters considered for the individual algorithms are reported in Sects. "NARX model architectures", "Multilayer Perceptron (MLP)" and "Random Forest (RF)". Rainfall was used as an exogenous input for the prediction of the streamflow. Furthermore, the time series were split with a 90–10% ratio for the training and testing stages, respectively. In preliminary tests, this subdivision proved to be optimal to guarantee high performance even in the prediction of flood peaks, while still preserving a sufficiently long testing period. Therefore, the period between January 1961 and March 2012 was considered for the training stage. Then, the subsequent period between April 2012 and December 2017 was considered for the testing stage. The Bayesian Optimization (BO) procedure was used for the selection of the ML hyperparameters and the optimal number of lagged values²⁹. In ML applications, the BO process aims to build a probability model of the objective function in order to select the most promising hyperparameters. For a detailed description of the BO procedure, please refer to the relevant literature³⁰.

NARX model architectures. NARX is a particular RNN generally used for time series modeling, made up of interconnected nodes that serve as artificial neurons, receiving one or more inputs and processing them via a nonlinear activation function to produce an output. The NARX model can be formulated as:

$$y(t) = f(y(t-1), y(t-2), \dots, y(t-f_d), x(t-1), x(t-2), \dots, x(t-p_d)) \quad (1)$$

where $x(t)$ and $y(t)$ indicate the exogenous input (i.e., precipitation) and the target (i.e., streamflow rate) at time t , respectively, p_d and f_d that represent the precipitation and flow rates lagged values, respectively. The NARX architecture consists of three layers (Fig. 3). The first is the input layer, which receives the input parameters. The second is the hidden layer, which represents the computational stage between input and output. The third is the output layer, which provides the predicted value. Then, the estimated output was fed back as input value for the iterative computation at the next instant³¹ (dashed line in Fig. 3). For the hidden layer, a sigmoid activation

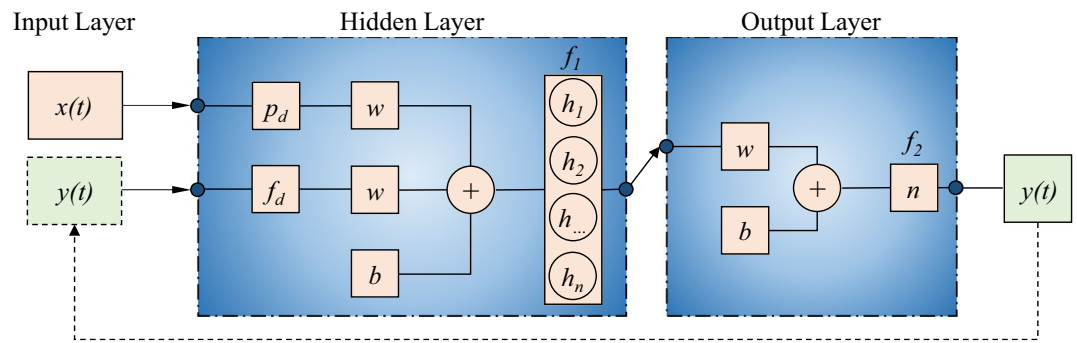


Figure 3. Sketch of the NARX architecture.

function f_1 was used, which is particularly suitable in neural networks trained through back-propagation algorithms. Moreover, the sigmoid function is derivable, making easier the neural network weights learning³². For the output layer, a linear activation function f_2 with one neuron n was used. Weight w and bias b were optimized by means of the Bayesian Regularization (BR) back-propagation training algorithm³³, which led to the best predictions compared with the other two training algorithms preliminarily tested, the Levenberg–Marquardt (LM) and the Scaled Conjugate Gradient (SCG). This agrees with previous literature studies that showed a slower convergence with, however, better performances for BR with respect to LM and SCG³⁴.

The BO procedure led to the optimal values of both optimal number of hidden nodes ($h_1, h_2, h_3, \dots, h_n$, in Fig. 3) and of p_d and f_d . The NARX process was stopped when one of the following conditions was met³⁵: maximum number of epochs, settled equal to 1000; LM adjustment parameter, settled equal to 1×10^{-10} ; error gradient below a minimal value, settled equal to 1×10^{-7} .

Multilayer Perceptron (MLP). MLP is a particular type of feedforward ANN^{36,37} with a similar structure to NARX, with three types of layers: input, hidden, and output (Fig. 4). The input layer is made up of a set of nodes corresponding to the input variables. One or more hidden layers contain neurons that process the values included in the input layer based on a weighted linear sum followed by a non-linear activation function. Then, the output layer gets the results from the last hidden layer, providing the expected values. Backpropagation learning algorithm was used for the training of the MLP neurons. The optimal structure of the MLP network for the present study includes one hidden layer, a neuron number equal to 10, and a Sigmoid activation function. Moreover, the optimal learning and momentum rates of the backpropagation algorithm were 0.3 and 0.2, respectively.

Random forest (RF). Random Forest (Fig. 5) is an ensemble of regression tree algorithms³⁸. Each tree is characterized by root and internal nodes which, respectively, include the training data and indicate the input variables conditions, and by leaves, which are the real values assigned to the target.

The development of a regression tree model consists of a recursive subdivision of the input data set into subsets, where predictions for each subset were achieved through a multivariable linear regression model. The growth of the trees is also an iterative procedure, where each subset is divided into small branches, assessing all the possible split for each field and finding, for each stage, the subdivision in two separate partitions that leads to the minimum squared deviation:

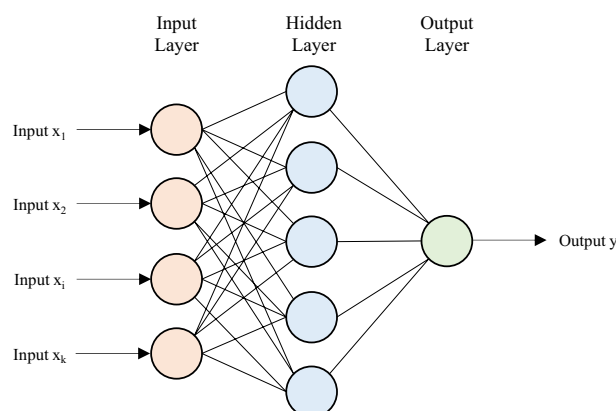


Figure 4. Sketch of the MLP architecture.

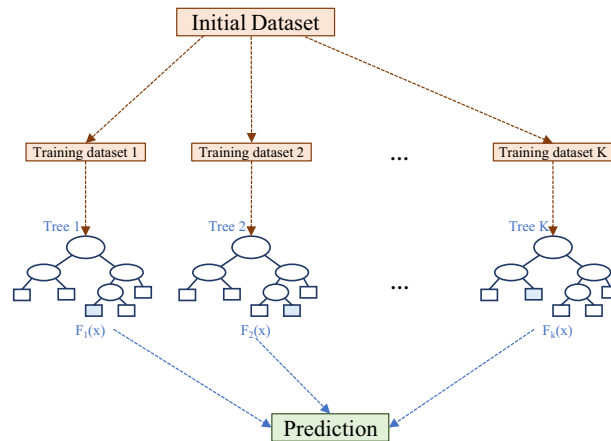


Figure 5. Sketch of the RF architecture.

$$R(t_{RF}) = \frac{1}{N(t)} \sum_{i \in t_{RF}} (y_i - y_m(t_{RF}))^2 \quad (2)$$

where $N(t)$ is the t_{RF} node's sample size, y_i is the target variable in the i^{th} unit, and y_m is the mean target variable in the node t_{RF} . $R(t_{RF})$ provides the “impurity” at each node. The algorithm stops when the minimum impurity is reached or based on when a different stopping rule is encountered. In addition, overfitting risk is reduced through a pruning process.

It should be noted that both MLP-RF and NARX-MLP-RF models were not particularly sensitive to the number of trees, which was set equal to 100 for all rivers and models.

Evaluation of model performance. The performance of the models was evaluated as the forecast horizon increased from 1 to 7 days ahead, based on five different evaluation metrics: the Coefficient of determination (R^2), RMSE, the Mean Absolute Error (MAE), the Mean Absolute Percentage Error (MAPE) and the Mean Directional Accuracy (MDA). A description of the evaluation metrics is reported in Table 2.

Results

Streamflow rate predictions on reference rivers. This section focuses primarily on flow forecasting in three reference rivers, chosen to evaluate the performance of different forecasting models in areas of the UK characterized by different rainfall regimes. The evaluation metrics for the training and testing stages, calculated for all rivers, forecasting models and temporal horizon, are shown in Tables 3, 4 and 5. In addition, Figures from 6 to 10 show the comparison between measured and predicted flow rate during the testing stage, for the different prediction models and forecast horizons.

Coefficient of determination Evaluates the goodness of fit in a regression model. It ranges between 0 (the model does not predict the outcome) to 1 (the model perfectly predicts the outcome)	$R^2 = 1 - \frac{\sum_{i=1}^n (Q_p^i - Q_A^i)^2}{\sum_{i=1}^n (\overline{Q_A} - Q_A^i)^2} \quad (3)$
Root Mean Square Error Root of total squared error between predicted and actual streamflow rate normalized by the number of samples. It ranges between 0 and $+\infty$ with lower values indicating more accurate models	$RMSE = \sqrt{\frac{\sum_{i=1}^n (Q_p^i - Q_A^i)^2}{s}} \quad (4)$
Mean Absolute Error Absolute error between the predicted and actual streamflow rate normalized by the number of samples. It ranges between 0 and $+\infty$ with lower values indicating more accurate models	$MAE = \frac{\sum_{i=1}^n Q_p^i - Q_A^i }{s} \quad (5)$
Mean Absolute Percentage Error Relative error between predicted and actual streamflow rate normalized by the number of samples It ranges between 0 and $+\infty$ with lower values indicating more accurate models	$MAPE = \frac{\sum_{i=1}^n \left \frac{Q_p^i - Q_A^i}{Q_A^i} \right }{s} \quad (6)$
Mean Directional Accuracy Compares predicted and actual direction (increasing or decreasing), providing the probability that the forecasting model can detect the correct direction along the time series. It ranges between 0 and 100%, with higher values indicating more accurate models	$MDA = \frac{\sum_{i=1}^n 1_{\text{sgn}(Q_A^i - Q_A^{i-1}) = \text{sgn}(Q_p^i - Q_A^{i-1})}}{s} \quad (7)$

Table 2. Evaluation metrics for NARX modeling. where Q_A^i = measured streamflow rate for the i^{th} data and Q_p^i = predicted streamflow rate for the i^{th} data, $\overline{Q_A}$ = mean streamflow rate, n = number of samples, $\text{sgn}(\cdot)$ = sign function, 1 = indicator function.

Time ahead	Stage	Metrics	Bure At Ingworth	Dee At Manley Hall	Dee At Woodend	Deveron At Muireisk	Leven At Newby Bridge	Nith At Friars Carse	Ribble At Samesbury	Severn At Montford	Spey At Boat O Brig	Tamar At Gunnislake	Tay At Ballathie	Teifi At Glanteifi	Test At Broadlands	Teviot At Ormiston Mill	Thames At Kingston	Trent At Colwick	Tweed at Peebles	Wye At Redbrook
1 day	Train	R ²	0.94	0.97	0.76	0.87	0.99	0.91	0.92	0.98	0.87	0.95	0.95	0.97	0.99	0.91	0.99	0.98	0.89	0.97
		RMSE (m ³ /s)	0.14	5.07	17.16	7.53	1.13	11.57	14.35	7.69	18.37	6.73	33.39	5.41	0.5	9.11	8.79	11.18	5.52	15.74
		MAE (m ³ /s)	0.06	2.62	8.09	2.59	0.63	5.25	6.68	4.07	9.44	2.88	17.77	2.52	0.23	3.94	5.51	5.89	2.15	7.63
		MAPE (%)	4.19	7.56	18.72	13.59	5.26	20.21	18.82	10.03	12.11	13.6	8.94	9.31	1.92	16.95	12.49	6.06	11.99	10.1
		MDA (%)	75.8	79.27	69.77	75.21	87.73	79.22	79.37	79.56	72.07	81.89	73.33	80.95	76.89	73.82	75.11	79.12	74.27	77.14
	Test	R ²	0.93	0.95	0.64	0.76	0.99	0.89	0.9	0.97	0.82	0.91	0.94	0.97	0.99	0.88	0.98	0.97	0.86	0.96
		RMSE (m ³ /s)	0.15	6.95	32.35	9.74	1.38	11.93	15.44	8.36	22.17	8.5	39.59	5.67	0.6	10.65	9.92	12.51	7.5	16.9
		MAE (m ³ /s)	0.06	2.79	9.57	3.04	0.74	6	7.28	4.13	9.59	3.16	19.91	2.8	0.28	4.44	5.96	6.22	2.77	7.81
		MAPE (%)	4.43	8.71	19.71	14.86	6.31	21.95	20.29	10.68	12.7	15.48	9.71	9.75	2.01	17.93	17.7	6.47	12.33	10.41
		MDA (%)	71.36	76.16	67.58	71.91	85.29	78.31	77.07	78.23	70.41	78.73	69.92	80.18	73.93	73.12	73.97	74.22	72.8	75.52
3 days	Train	R ²	0.94	0.97	0.71	0.86	0.99	0.9	0.9	0.98	0.82	0.95	0.95	0.97	0.99	0.9	0.99	0.98	0.89	0.97
		RMSE (m ³ /s)	0.15	5.24	18.89	7.56	1.24	11.88	14.86	7.81	21.99	7.16	34.05	5.64	0.53	11.18	8.83	11.69	5.55	15.77
		MAE (m ³ /s)	0.06	2.86	8.19	2.86	0.66	5.49	7.03	4.34	9.67	3.35	17.91	2.66	0.23	3.94	5.63	6.4	2.19	7.84
		MAPE (%)	4.25	8.32	19.33	14.83	5.48	23.09	20.36	11.12	12.41	15.53	9.24	9.71	1.93	20.13	12.69	6.43	12.3	10.78
		MDA (%)	75.71	78.62	68.78	73.43	87.38	76.79	76.74	78.64	71.45	80.26	72.29	79.96	75.26	72.84	72.67	79.12	73.38	75.21
	Test	R ²	0.92	0.94	0.45	0.61	0.99	0.88	0.89	0.97	0.79	0.89	0.94	0.96	0.99	0.79	0.98	0.97	0.83	0.95
		RMSE (m ³ /s)	0.17	7.72	40.32	12.59	1.45	12.24	16.75	8.4	24.39	9.83	40.27	5.97	0.61	12.3	10.2	12.63	8.54	17.03
		MAE (m ³ /s)	0.07	2.96	10.52	3.26	0.78	6.34	7.4	4.38	9.81	3.51	20.94	2.96	0.3	4.71	6.08	6.51	2.8	8.01
		MAPE (%)	4.64	9.41	20	15.84	6.36	23.75	23.39	11.23	12.75	17.9	9.73	10.21	2.16	20.23	18.9	6.68	13.12	10.91
		MDA (%)	71.28	75.58	67.06	71.21	85.25	76.16	75.53	77.24	69.77	77.74	69.83	79.64	73.53	71.82	72.61	73.9	70.5	75.03
7 days	Train	R ²	0.92	0.95	0.58	0.79	0.98	0.79	0.77	0.98	0.82	0.9	0.93	0.95	0.99	0.76	0.99	0.98	0.89	0.97
		RMSE (m ³ /s)	0.16	6.8	24.74	10.09	1.95	17.11	22.01	8.27	22.28	9.52	38.58	7.04	0.53	12.57	9.44	12.61	7.48	16.63
		MAE (m ³ /s)	0.06	3.17	8.73	3.03	0.75	7.78	9.8	4.8	10.18	4.02	19.52	3.18	0.26	5.26	5.97	7.21	2.98	8.66
		MAPE (%)	4.61	9.13	20.84	16.93	6.6	25.44	24.92	13.67	12.85	18.72	10.23	10.73	2.19	22.74	13.57	7.5	15.5	13.2
		MDA (%)	75.01	76.4	68.04	70.01	86.49	75.21	74.86	77.31	71.44	78.43	72.19	77.95	74.32	71.99	72.66	78.28	72.09	73.58
	Test	R ²	0.88	0.87	0.43	0.57	0.97	0.75	0.77	0.97	0.66	0.86	0.92	0.94	0.99	0.63	0.98	0.97	0.8	0.95
		RMSE (m ³ /s)	0.21	11.94	67.37	15.52	2.23	18.19	25.5	8.69	35.99	10.91	45.96	7.88	0.71	21.81	10.6	13.6	9.53	17.38
		MAE (m ³ /s)	0.07	3.21	10.91	3.65	0.84	9.06	10.57	4.95	11.24	4.4	22.76	3.53	0.33	6.84	6.33	7.22	3.52	8.88
		MAPE (%)	5.06	10.96	21.55	17.55	8.13	26.87	27.57	13.89	13.73	23.81	10.38	11.48	2.37	23.87	19.4	7.73	16.02	13.37
		MDA (%)	70.75	75.21	66.76	68.67	84.44	72.95	74.53	75.31	69.12	76.73	69.23	77.93	73.51	70.26	72.59	69.33	72.89	

Table 3. Evaluation metrics for NARX modeling.

Time ahead	Stage	Metrics	Bure At Ingworth	Dee At Manley Hall	Dee At Woodend	Deveron At Muireisk	Leven At Newby Bridge	Nith At Friars Carse	Ribble At Samesbury	Severn At Montford	Spey At Boat O Brig	Tamar At Gunnislake	Tay At Ballathie	Teifi At Glanteifi	Test At Broadlands	Teviot At Ormiston Mill	Thames At Kingston	Trent At Colwick	Tweed at Peebles	Wye At Redbrook
1 day	Train	R ²	0.9	0.95	0.83	0.83	0.98	0.88	0.85	0.95	0.86	0.9	0.93	0.94	0.98	0.95	0.98	0.96	0.89	0.95
		RMSE (m ³ /s)	0.19	6.98	14.78	8.06	2.08	12.15	18.34	11.22	19.54	9.34	36.32	7.47	0.65	6.62	10.58	15.79	5.43	20.02
		MAE (m ³ /s)	0.07	3.29	7.19	3.16	0.97	5.89	9.6	6.12	9.37	4.04	17.92	3.23	0.28	2.81	6.07	8.32	2.24	9.47
		MAPE (%)	4.75	8.25	15.96	14.44	7.21	19.6	30.21	13.78	11.43	17.31	9.07	10.21	2.06	11.24	12.37	8.09	11.99	10.32
		MDA (%)	71.34	71.5	67.94	70.56	80.01	76.63	73.96	70.6	74.42	74.91	71.48	76.2	75.21	87.28	81.3	73.65	76.05	75.88
	Test	R ²	0.88	0.94	0.76	0.81	0.96	0.87	0.85	0.93	0.85	0.9	0.91	0.94	0.98	0.94	0.97	0.94	0.89	0.94
		RMSE (m ³ /s)	0.2	7.67	23.59	8.79	3.1	14.34	21.78	13.02	19.8	9.89	51.31	7.97	0.76	7.99	11.89	17.59	7.44	20.06
		MAE (m ³ /s)	0.08	3.29	9.76	3.17	1.36	8.14	10.38	6.54	10.14	4.83	23.3	3.67	0.32	3.31	6.55	8.99	3.32	10.25
		MAPE (%)	5.16	9.06	19.47	14.71	7.59	22.31	32.48	13.83	13.04	17.43	9.63	10.5	2.2	12.76	12.5	8.74	13.42	11.45
		MDA (%)	71.05	71.14	67.24	68.62	77.59	76.6	73.13	69.32	73.53	74.79	71.05	75.44	70.23	86.32	80.63	72.19	76.05	74.42
3 days	Train	R ²	0.88	0.93	0.76	0.8	0.92	0.87	0.85	0.86	0.81	0.88	0.9	0.93	0.97	0.91	0.97	0.85	0.89	0.87
		RMSE (m ³ /s)	0.21	8.45	18.83	9.06	3.79	12.46	18.58	19.3	23.61	9.86	43.82	8.58	0.88	6.64	15.27	30.91	5.48	32.27
		MAE (m ³ /s)	0.1	4.88	10.92	4.26	2.09	7.16	9.76	11.47	14.42	4.97	24.82	4.55	0.47	3.4	8.72	16.34	2.87	17.57
		MAPE (%)	7.48	15.12	30.61	25.76	16.96	38.5	33.26	30.02	22.08	29.64	14.71	18.37	3.64	18.86	19.03	15.32	20.7	23.51
		MDA (%)	70.88	70.86	67	69.42	72.49	76.2	73.03	68.98	71.85	71.02	69.77	75.51	74.62	75.23	73.1	70.51	75.95	66.53
	Test	R ²	0.86	0.9	0.72	0.78	0.89	0.84	0.85	0.84	0.77	0.87	0.9	0.92	0.96	0.89	0.95	0.83	0.86	0.84
		RMSE (m ³ /s)	0.22	9.77	24.49	9.06	5.23	15.87	21.88	20.51	26.3	11.23	55.29	8.8	0.98	9.04	15.4	31.18	8.46	32.7
		MAE (m ³ /s)	0.11	5.28	12.41	4.54	2.73	8.39	10.59	11.65	14.65	5.28	30.63	4.95	0.51	4.46	8.85	16.59	3.66	18.6
		MAPE (%)	8.33	19.12	32.36	25.99	17.49	43.04	40.16	30.22	23.4	35.89	15.63	18.71	3.93	21.13	30.1	16.53	20.93	25.28
		MDA (%)	69.77	70.16	66.36	68.41	72.33	75.4	72.09	66.45	70.83	70.81	69.32	74.67	69.79	75.16	71.78	70.15	74.99	66.3
7 days	Train	R ²	0.81	0.91	0.73	0.74	0.88	0.85	0.82	0.79	0.8	0.87	0.89	0.92	0.96	0.87	0.93	0.79	0.87	0.83
		RMSE (m ³ /s)	0.27	9.48	21.05	10.4	4.91	13.67	20.17	24.01	25.01	11.29	48.68	9.49	1.11	8.19	25.01	35.72	8.12	38.42
		MAE (m ³ /s)	0.15	6.4	14.36	6.28	3.24	9.68	10.16	16.07	16.74	7.75	32.31	6.16	0.71	4.96	14.83	20.93	3.51	24.3
		MAPE (%)	11.55	24.92	49.79	47.37	29.99	71.63	35.6	54.7	28.08	71.93	23.63	34.13	6.16	34.85	30.33	22.14	23.2	42.77
		MDA (%)	69.77	70.75	66.8	66.76	70.76	74.38	64.81	66.92	70.04	69.06	66.09	73.24	72.84	75.2	71.2	70.41	70.4	65
	Test	R ²	0.78	0.89	0.69	0.73	0.87	0.81	0.81	0.77	0.75	0.85	0.88	0.91	0.95	0.85	0.88	0.78	0.76	0.78
		RMSE (m ³ /s)	0.27	11.02	26.26	10.75	5.98	17.22	23.43	24.84	27.9	11.34	63.14	9.85	1.19	10.78	25.08	37.14	8.62	38.68
		MAE (m ³ /s)	0.15	7.42	15.56	6.37	3.66	11.03	10.93	16.36	16.85	8.06	41.27	6.77	0.76	5.79	15.1	21.44	4.37	25.05
		MAPE (%)	12.32	36.3	50.18	47.94	36.55	74.04	45.98	55.51	30.44	73.99	26.65	35.25	6.77	34.86	23.9	25.2	24.9	44.44
		MDA (%)	68.9	69.18	65.73	64.97	69.8	70.66	62.84	64.47	66.8	67.44	65.86	72.44	69.4	74.02	70.31	67.75	68.73	64.18

Table 4. Evaluation metrics for MLP-RF modeling.

Time ahead		Stage	Metrics	Bure At Ingworth	Dee At Manley Hall	Dee At Woodend	Deveron At Muireisk	Leven At Newby Bridge	Nith At Friars Carse	Ribble At Samesbury	Severn At Montford	Spey At Boat O Brig	Tamar At Gunnislake	Tay At Ballathie	Teifi At Glanteifi	Test At Broadlands	Teviot At Ormiston Mill	Thames At Kingston	Trent At Colwick	Tweed at Peebles	Wye At Redbrook
1 day	Train	R ²	0.94	0.98	0.84	0.89	0.99	0.92	0.92	0.98	0.88	0.95	0.95	0.97	0.99	0.95	0.99	0.98	0.91	0.97	
		RMSE (m ³ /s)	0.13	4.99	14.33	6.63	1.05	10.94	13.91	7.08	17.54	6.64	32.42	5.29	0.5	6.6	8.35	10.53	5.23	15.65	
		MAE (m ³ /s)	0.06	2.6	7.02	2.58	0.56	5.15	6.54	3.94	9.31	2.88	17.29	2.51	0.23	2.62	5.28	5.75	2.1	7.43	
		MAPE (%)	4.08	7.37	13.28	13.28	4.83	19.31	18.46	9.76	11.41	12.73	8.84	8.64	1.87	11.12	11.96	5.79	11.92	9.55	
		MDA (%)	76.94	80.26	70.21	75.61	87.93	80.41	80.11	79.94	74.63	83.08	74.17	82.38	78.23	87.63	81.61	80.31	76.99	77.14	
	Test	R ²	0.94	0.96	0.77	0.82	0.99	0.9	0.91	0.97	0.85	0.93	0.94	0.97	0.99	0.94	0.98	0.97	0.9	0.96	
		RMSE (m ³ /s)	0.14	6.47	23.12	8.63	1.14	11.46	14.87	8.22	19.75	7.8	38.89	5.54	0.58	7.88	9.75	12.25	6.87	16.4	
		MAE (m ³ /s)	0.06	2.78	9.34	3.02	0.64	5.94	7.11	4.13	9.59	3.14	19.88	2.76	0.27	3.18	5.8	6.21	2.71	7.46	
		MAPE (%)	4.37	8.7	19.42	14.68	6.01	20.42	20.25	10.06	12.38	14.14	9.5	9.43	1.94	12.48	12.5	6.36	12.28	9.65	
		MDA (%)	73.33	77.17	68.48	73.28	85.7	79.47	78.12	78.77	73.64	80	71.15	81.14	75.36	86.5	80.78	75.5	76.12	75.95	
3 days	Train	R ²	0.94	0.97	0.77	0.88	0.99	0.91	0.91	0.98	0.84	0.95	0.95	0.97	0.99	0.91	0.99	0.98	0.9	0.97	
		RMSE (m ³ /s)	0.14	5.17	17.54	7.05	1.06	11.29	14.35	7.74	20.48	6.8	33.21	5.45	0.51	6.53	8.55	11.51	5.29	15.77	
		MAE (m ³ /s)	0.06	2.77	8.18	2.72	0.57	5.26	6.89	4.1	9.55	3.03	17.81	2.52	0.23	3.25	5.39	6.2	2.17	7.8	
		MAPE (%)	4.16	7.7	19.3	14.37	5.04	19.3	20.07	10.09	12.32	12.85	9.09	8.79	1.93	17.89	12.31	6.36	12.23	10.3	
		MDA (%)	76.6	79.27	69.42	73.82	87.53	79.42	77.44	78.84	72.74	80.9	74.07	81.05	76.45	75.32	74.56	79.81	75.99	75.98	
	Test	R ²	0.93	0.95	0.73	0.79	0.99	0.89	0.9	0.97	0.81	0.91	0.94	0.97	0.99	0.89	0.98	0.97	0.86	0.96	
		RMSE (m ³ /s)	0.16	7.12	24.15	9.02	1.33	11.68	16.14	8.25	22.7	8.88	39.64	5.92	0.6	9.01	9.81	12.6	7.67	16.75	
		MAE (m ³ /s)	0.06	2.94	10.33	3.12	0.65	6.05	7.31	4.16	9.6	3.22	20.16	2.76	0.28	4.34	5.85	6.48	2.79	7.89	
		MAPE (%)	4.54	8.88	19.77	15.78	6.24	21.06	22.72	10.43	12.59	14.69	9.57	9.82	2.05	18.78	18.8	6.67	12.61	10.72	
		MDA (%)	72.64	76.56	68.18	71.37	85.48	78.4	76.75	77.49	70.91	78.57	71.06	80.87	75.32	75.28	74.52	75.08	75.04	75.27	
7 days	Train	R ²	0.93	0.96	0.74	0.82	0.99	0.85	0.82	0.98	0.84	0.92	0.94	0.95	0.99	0.87	0.99	0.98	0.9	0.97	
		RMSE (m ³ /s)	0.15	6.26	20.98	9.06	1.78	13.33	19.85	7.78	20.72	8.62	35.84	6.43	0.52	8.07	8.99	11.58	6.7	16.18	
		MAE (m ³ /s)	0.06	2.94	8.36	2.91	0.72	6.97	8.86	4.26	9.62	3.63	18.14	2.91	0.24	4.85	5.68	6.34	2.7	7.82	
		MAPE (%)	4.6	8.46	20.27	15.72	5.97	24.8	22.62	10.75	12.55	17.29	9.97	10.44	2	20.62	12.73	6.41	13.88	11.06	
		MDA (%)	76.05	77.24	69.37	71.7	86.84	77.73	76.45	78.84	72.58	80.06	73.03	79.54	75.41	75.21	73.19	78.67	74.37	74.52	
	Test	R ²	0.9	0.9	0.7	0.73	0.98	0.81	0.81	0.97	0.75	0.88	0.93	0.95	0.99	0.85	0.98	0.97	0.84	0.96	
		RMSE (m ³ /s)	0.19	10.28	26.13	10.72	1.99	16.52	22.95	8.68	27.83	9.89	43.45	7.41	0.67	10.64	10.36	13.25	8.52	17.13	
		MAE (m ³ /s)	0.07	3.11	10.48	3.39	0.83	8.2	9.56	4.27	10.64	3.99	21.44	3.28	0.3	5.66	6.17	6.51	3.2	8.12	
		MAPE (%)	4.81	9.47	21.34	17.35	6.98	26.63	25.32	11.41	12.97	20.03	10.33	11.06	2.23	21.09	19.2	6.76	14.36	11.56	
		MDA (%)	72.59	76.21	67.75	70.25	85.3	75.24	76.31	77.44	70.01	78.24	70.98	78.97	74.43	74.33	73.03	75.06	71.73	73.28	

Table 5. Evaluation metrics for NARX-MLP-RF modeling.

The first river considered was Tay at Ballathie, Scotland, with the second highest average annual precipitation over the catchment area and the highest average annual flow rate among the 18 rivers analyzed (see Section “[Case studies and dataset](#)”). The NARX-MLP-RF hybrid model outperformed both NARX and MLP-RF models. The best performance was observed for the shortest forecast horizon $t = 1$ day, with the NARX model outperforming MLP-RF model for both training and testing stages. As can be seen in Fig. 6, NARX led to a more accurate prediction of the peak flow rates. However, compared to MLP-RF, NARX showed a tendency to overestimate the flow rates more frequently than MLP-RF. Therefore, the NARX-MLP-RF hybrid model, combined the advantages of both models, leading to more robust predictions compared with the two individual NARX and MLP-RF models. As the forecast horizon increases, a decrease in accuracy was observed for all models. Specifically, for $t = 3$ days (Fig. 7), the difference in prediction accuracy between the NARX and MLP-RF models is more marked, with the latter still showing a good ability to predict flow rate trends but with a more accentuated underestimation of the peaks, compared to $t = 1$ day. However, again the NARX-MLP-RF hybrid model resulted in the best forecasts, although metrics were only slightly better than the individual NARX model. The worst predictions were observed for $t = 7$ days (Fig. 8), with NARX showing a significant over- and underestimation of flow rates compared to shorter forecast horizons. Also, MLP-RF shows a decrease in performance with, however, a lower dispersion compared to NARX, particularly for the medium–low values of flow rate (Fig. 8b and d). Consequently, the best prediction was obtained with the NARX-MLP-RF hybrid model, which showed a limited accuracy reduction from a 3-day to 7-day ahead forecast horizon.

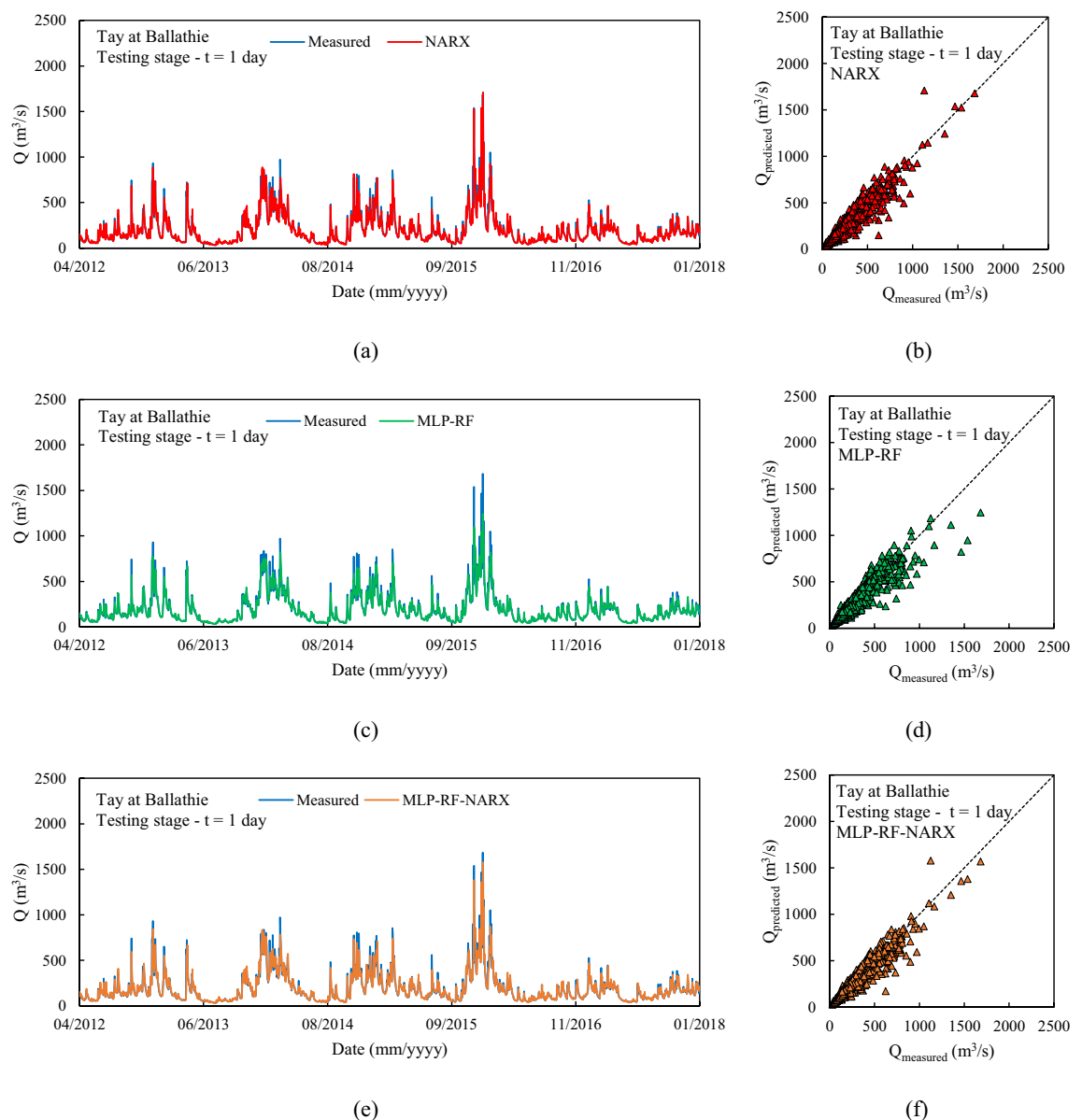


Figure 6. 1-day ahead predictions for Tay at Ballathie: NARX (a, b); MLP-RF (c, d); MLP-RF-NARX (e, f).

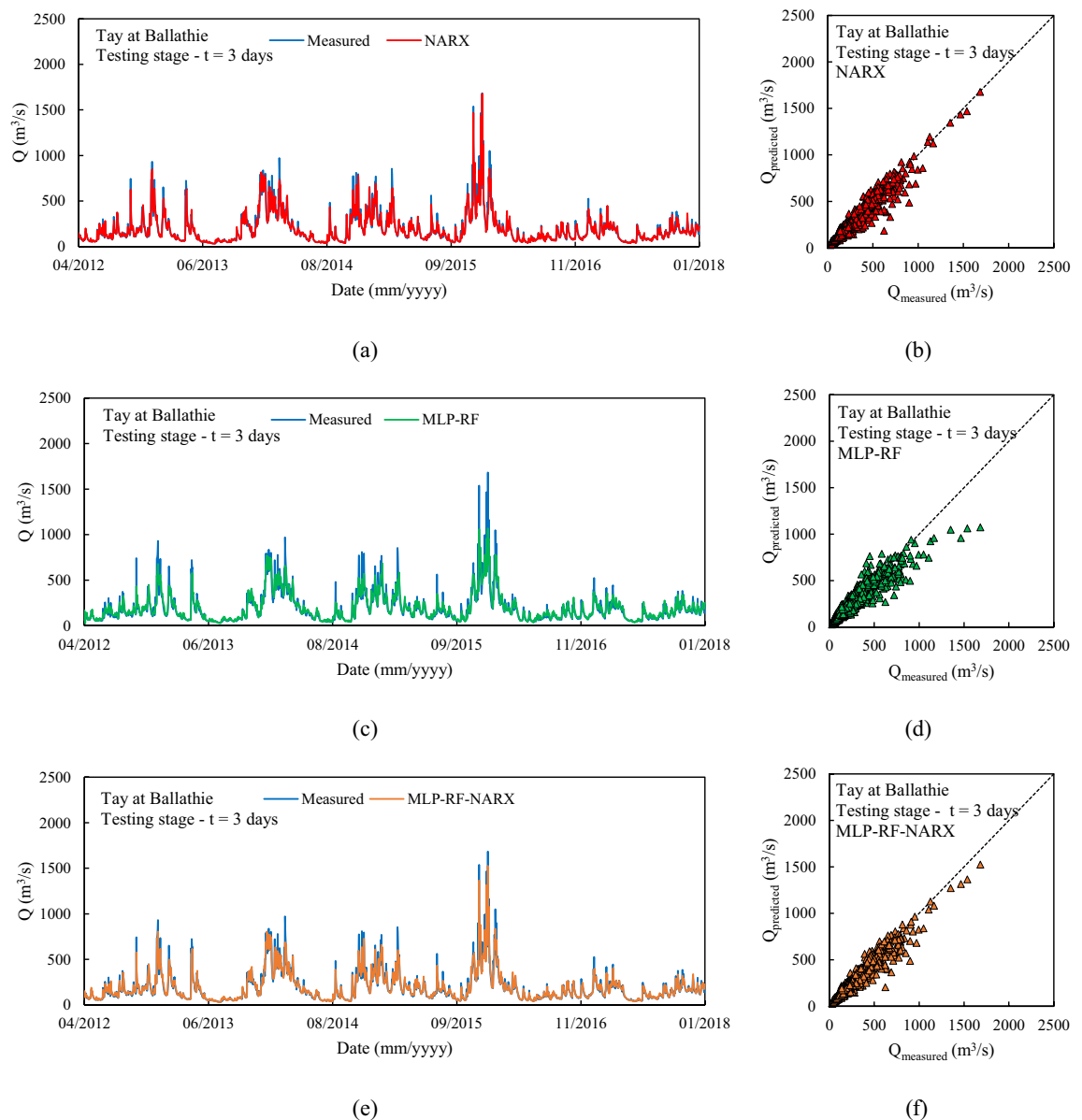


Figure 7. 3-days ahead predictions for Tay at Ballathie: NARX (a, b); MLP-RF (c, d); MLP-RF-NARX (e, f).

The second river analyzed in detail is the Ribble in Samlesbury, England. It showed, during the spring, a marked decreasing trend in both precipitation over the catchment area and streamflow. Figure 9 shows the comparison between measured and predicted flow rate, for forecast horizons of 1 day and 7 days, and for the NARX-MLP-RF hybrid model. Furthermore, the results for the individual models are shown in Tables 3 and 4. As for the testing stage, the best predictions were obtained for a forecast horizon of 1 day with the NARX-MLP-RF hybrid model, with $R^2 = 0.91$. The NARX model ($R^2 = 0.90$) resulted in slightly worse prediction than the hybrid model, while still providing more accurate forecasts than the MLP-RF model ($R^2 = 0.85$). Again, as the forecast horizon increases, a reduction of the prediction accuracy was observed for the three different models. However, for $t = 7$ days, MLP-RF ($R^2 = 0.81$) outperformed NARX ($R^2 = 0.77$), which, however, still led to higher MDA values, indicating a better ability to follow the flow rate trend (MLP-RF-MDA = 62.84%, NARX-MDA = 74.53%), whereas the NARX-MLP-RF hybrid model combined the strengths of the individual models leading to better predictions ($R^2 = 0.81$ and MDA = 76.31%).

The third reference river was the Thames at Kingston, in the south of England, which has the largest catchment area among the 18 rivers. This case study shows overall very accurate predictions for the three different forecast models and horizons. For $t = 1$ day and for the testing stage, R^2 values of up to 0.98 were calculated for MLP-RF and up to 0.99 for both NARX and the NARX-MLP-RF hybrid. The predictions became less accurate as the forecast horizon increased while maintaining higher accuracy under the same conditions, compared to the two previously investigated cases, with R^2 values up to 0.95 for MLP-RF and 0.98 for both NARX and NARX-MLP-RF,

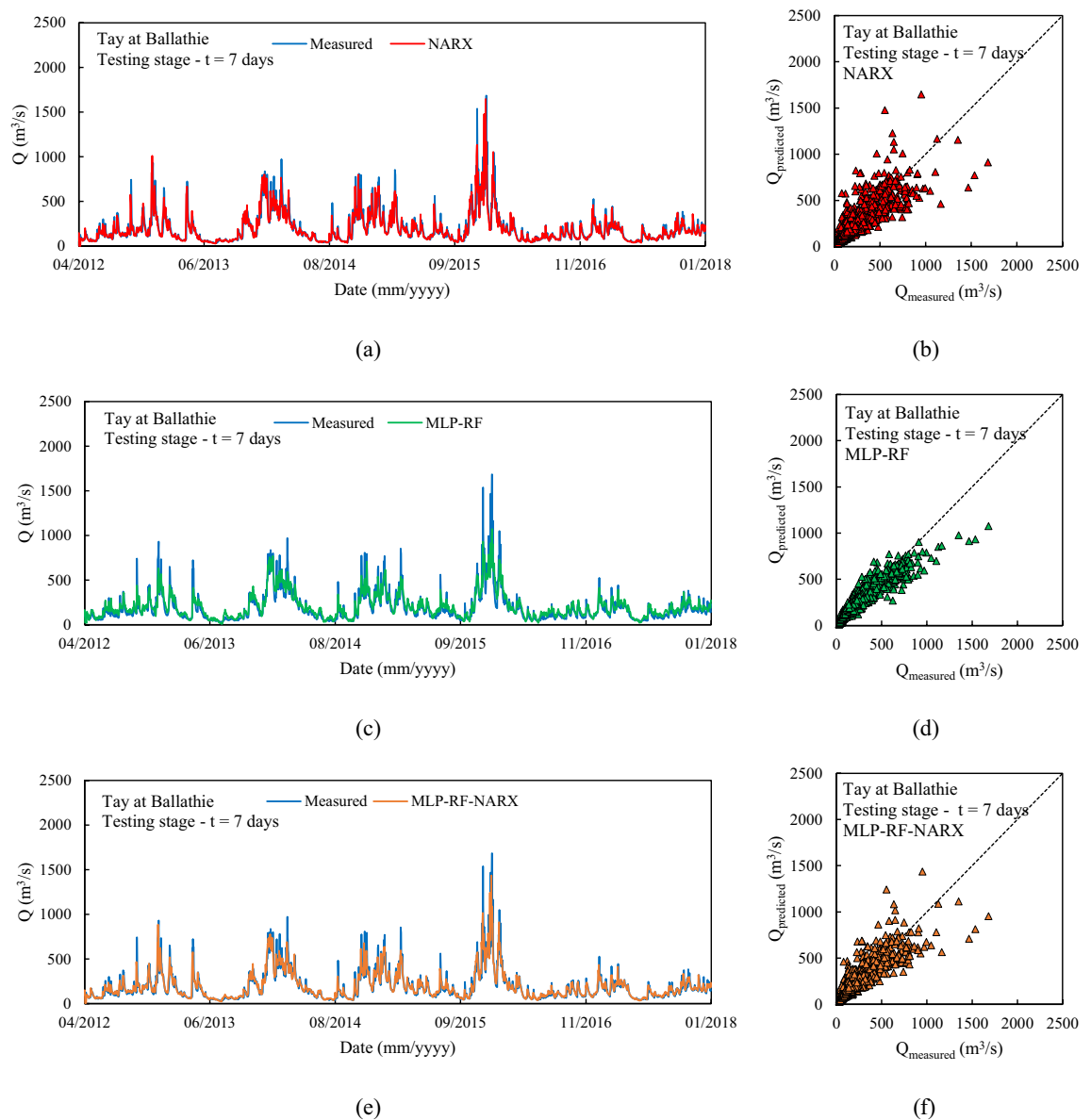


Figure 8. 7-days ahead predictions for Tay at Ballathie: NARX (a, b); MLP-RF (c, d); MLP-RF-NARX (e, f).

for $t = 3$ days. A marked decrease was observed only for $t = 7$ days for MLP-RF with $R^2 = 0.88$. Both NARX and NARX-MLP-RF showed an R^2 equal to 0.98, with a limited reduction in the other metrics (Fig. 10).

Overall, the high performance of the forecast models for the Thames at Kingston can be justified by particularly gradual variations in the flow rates, which facilitate the predictions of peaks along the time series, linked to the large catchment area and lower average rainfall compared to the rest of England, and with a homogeneous distribution throughout the year. These factors make the hybridization of NARX and MLP-RF less relevant in terms of forecast improvement. Conversely, forecast models for rivers with smaller catchments and higher but less homogeneous rainfall throughout the year, as in the case of Ribble at Samlesbury, benefited more from hybridization, with better forecasts and a lower reduction in performance as the forecast horizon increases.

One aspect investigated with special emphasis is the highest flow rates, which can represent critical scenarios as they can lead to flooding. From this point of view, relative errors were calculated with reference to the first decile of flow rates for the three different models and for different forecast horizons. The relative errors were calculated as the difference between the predicted and measured values, divided by the measured values. Histograms with the frequency of the relative errors for the three reference rivers are shown in Figs. 11, 12 and 13, respectively. For the Tay River at Ballathie (Fig. 11) and $t = 1$ day, the relative errors were in the range $-0.5 \div 0.4$, with an almost symmetrical distribution for all three models. In particular, the NARX-MLP-RF ensemble model showed the highest frequency of low relative errors, equal to 24% and 29% for relative errors between -0.1 and 0 and between 0 and 0.1 , respectively. MLP-RF, on the other hand, showed a lower frequency of relative errors between -0.1 and 0 and between 0 and 0.1 , amounting to 19% and 23%, respectively. The NARX model showed a

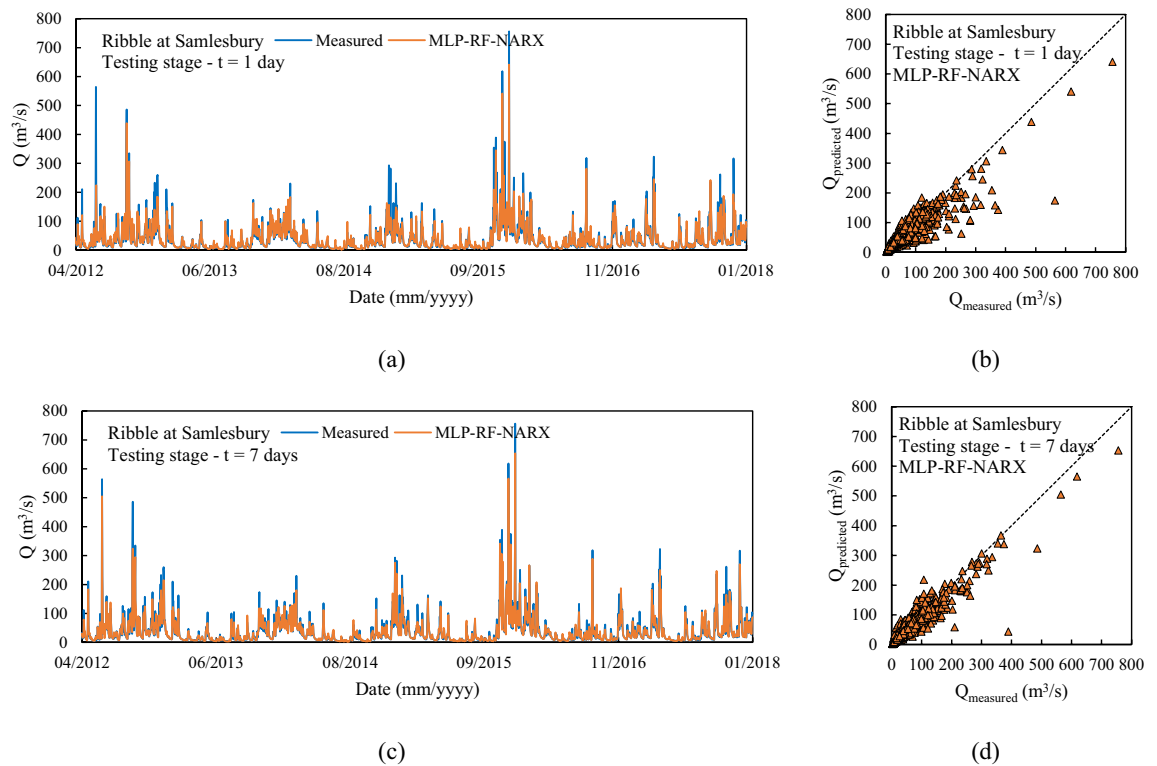


Figure 9. Predictions for Ribble at Samlesbury with MLP-RF-NARX model: $t = 1$ day (a, b); $t = 7$ days (c, d).

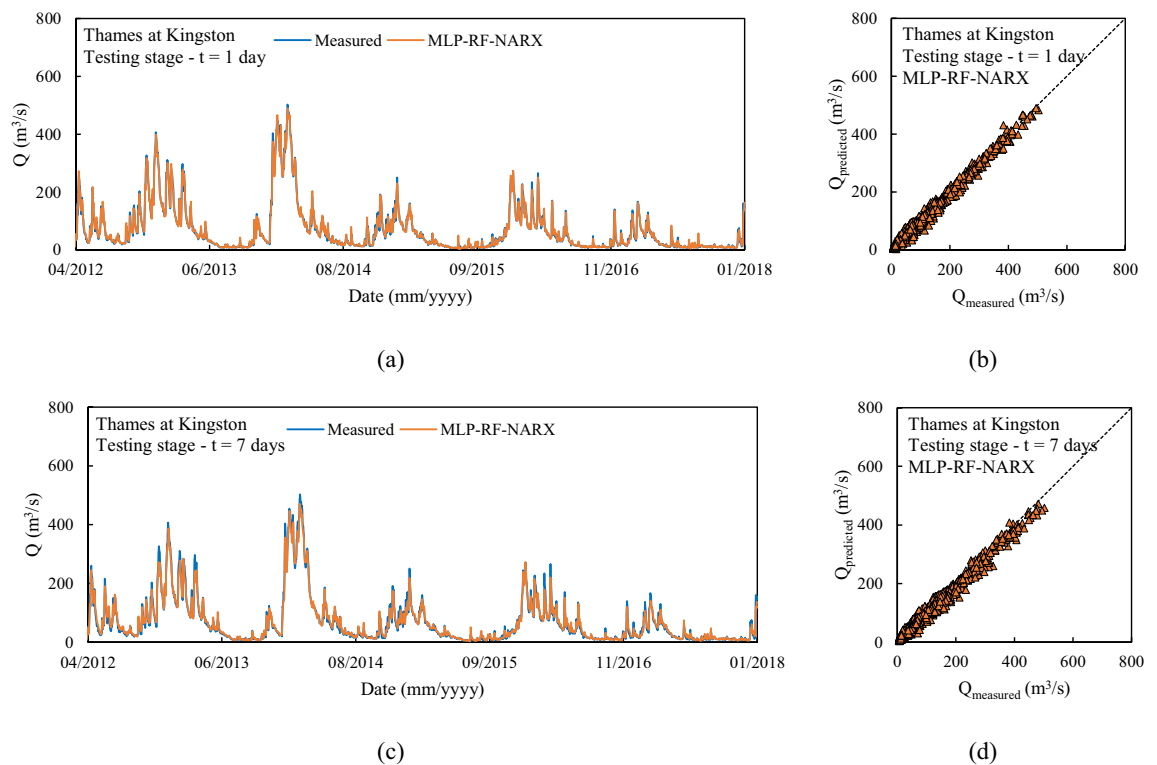


Figure 10. Predictions for the Thames at Kingston with MLP-RF-NARX model: $t = 1$ day (a, b); $t = 7$ days (c, d).

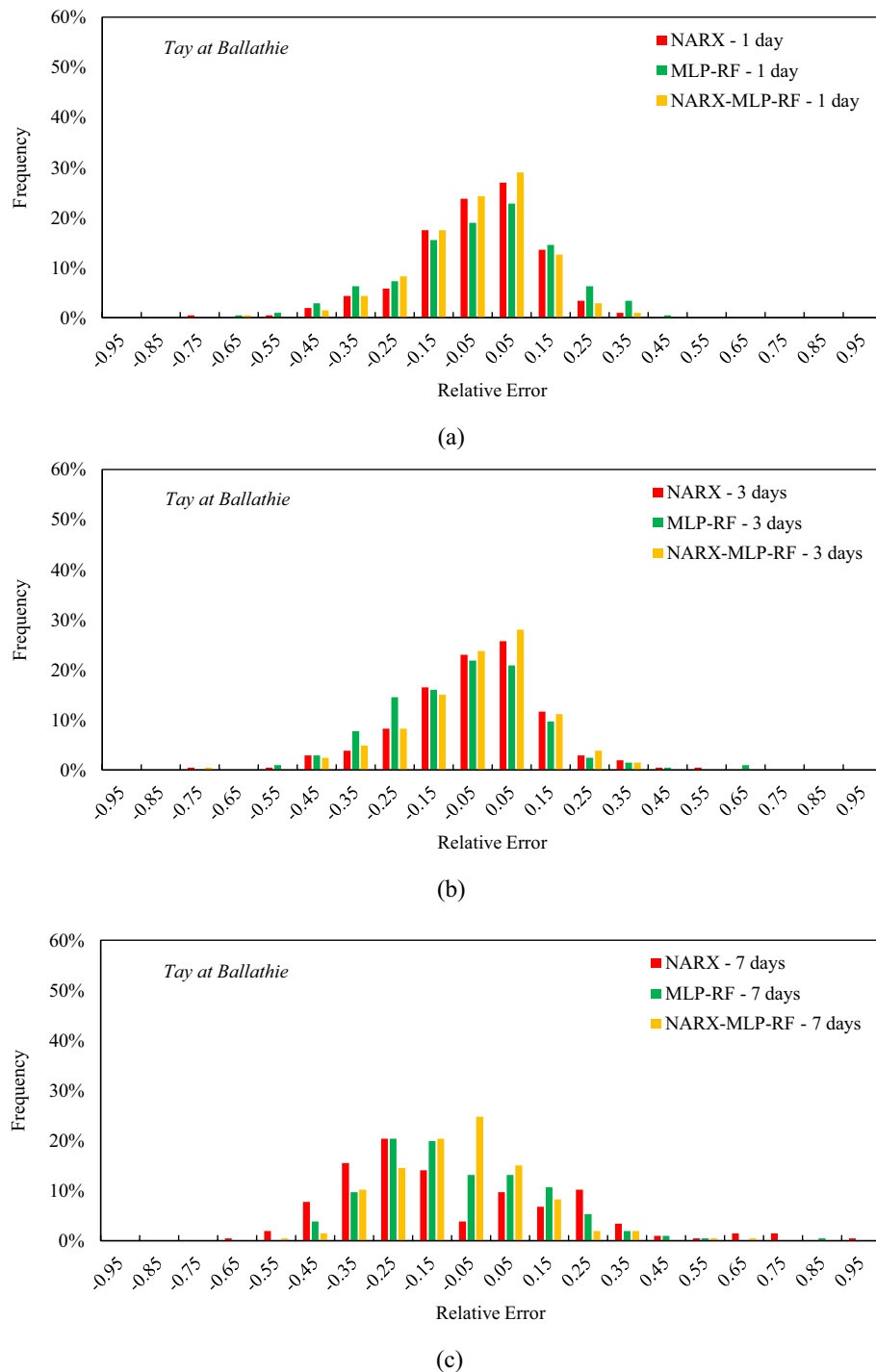


Figure 11. Frequency of the relative error for the tenth decile for Tay at Ballathie: $t = 1$ day (a); $t = 3$ days (b); $t = 7$ days (c).

similar frequency distribution to the NARX-MLP-RF ensemble model with, however, slightly lower frequencies for lower relative errors. As the forecast horizon increases, the accuracy of the three models is reduced. Thus, a decrease in frequency was observed for the lower relative errors, with a subsequent increase in frequency for the higher relative errors. For $t = 7$ days, the NARX-MLP-RF ensemble showed the highest frequency for the relative errors between -0.1 and 0 , i.e., 25%, maintaining a rather symmetrical distribution. In contrast, the NARX model showed a less symmetrical distribution with a frequency of around 20%, for relative errors between -0.3 and -0.2 . Frequencies in the order of 20% were also observed for the MLP-RF model, both for relative errors between -0.3 and -0.2 (as for NARX) and between -0.2 and -0.1 . This result showed a tendency for the NARX and MLP-RF models to underestimate peak flow rates.

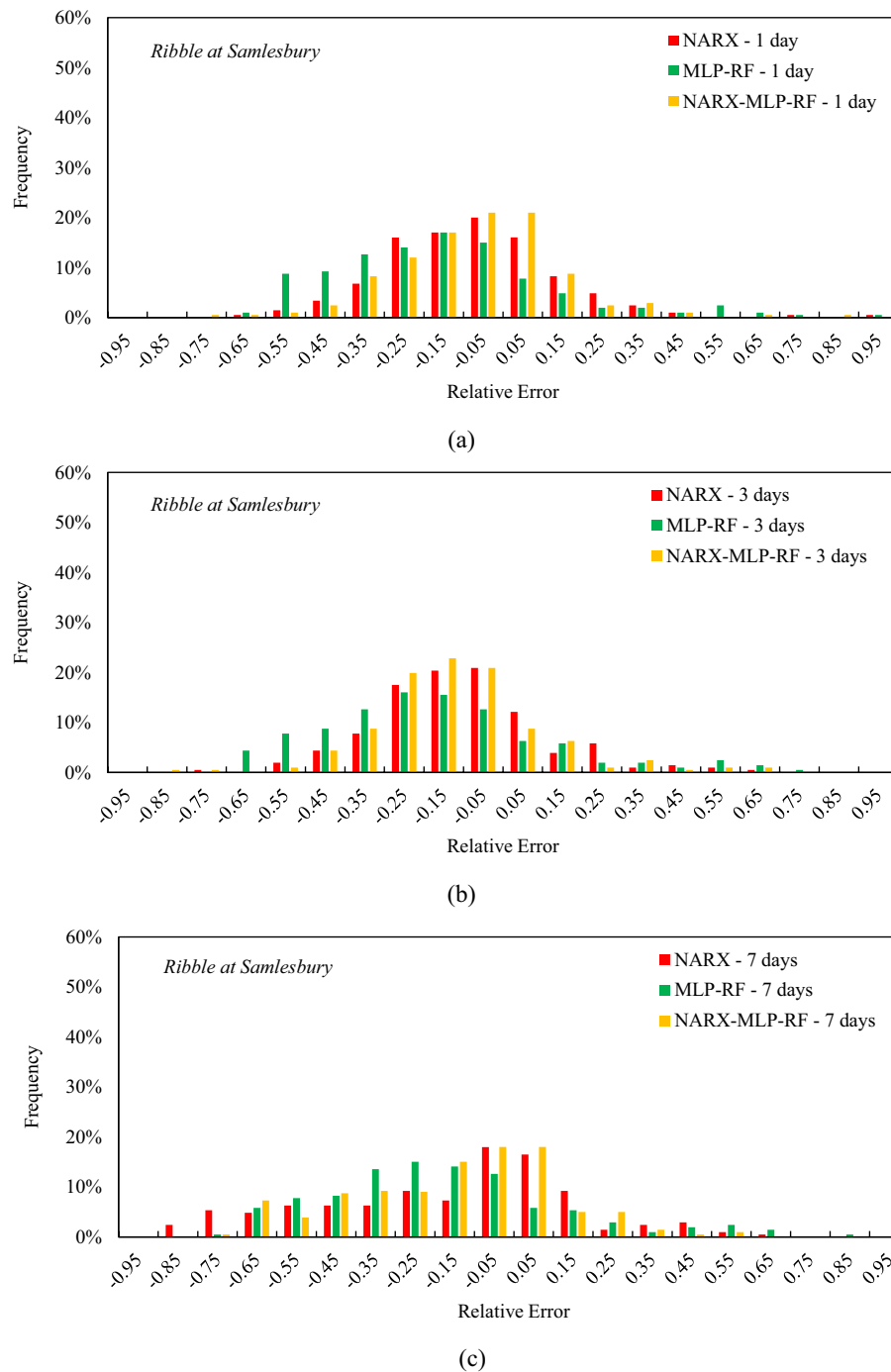


Figure 12. Frequency of the relative error for the tenth decile for Ribble at Samlesbury: $t = 1$ day (a); $t = 3$ days (b); $t = 7$ days (c).

For the Ribble at Samlesbury (Fig. 12) and $t = 1$ day, the relative errors were in the range -0.6 – 0.6 . The NARX-MLP-RF ensemble showed the highest frequency of low relative errors of 21% for both relative errors between -0.1 and 0 and between 0 and 0.1 , showing an almost symmetrical distribution. In contrast, MLP-RF showed a lower frequency of relative errors between -0.1 and 0 and between 0 and 0.1 . The latter also showed a peak frequency of 17% for relative errors between -0.2 and -0.1 , showing a more skewed distribution than the NARX-MLP-RF ensemble model. The NARX model showed lower frequencies, compared to NARX-MLP-RF, for the relative errors between -0.1 and 0 and between 0 and 0.1 , amounting to 20% and 16% respectively. As the prediction horizon increased, an increase in the variance of the relative error distributions was observed, with a reduction in the frequencies corresponding to the lowest relative errors. In particular, the NARX model also showed relative errors in the range between -0.9 and -0.8 , but with a very low frequency of 2%. All three models

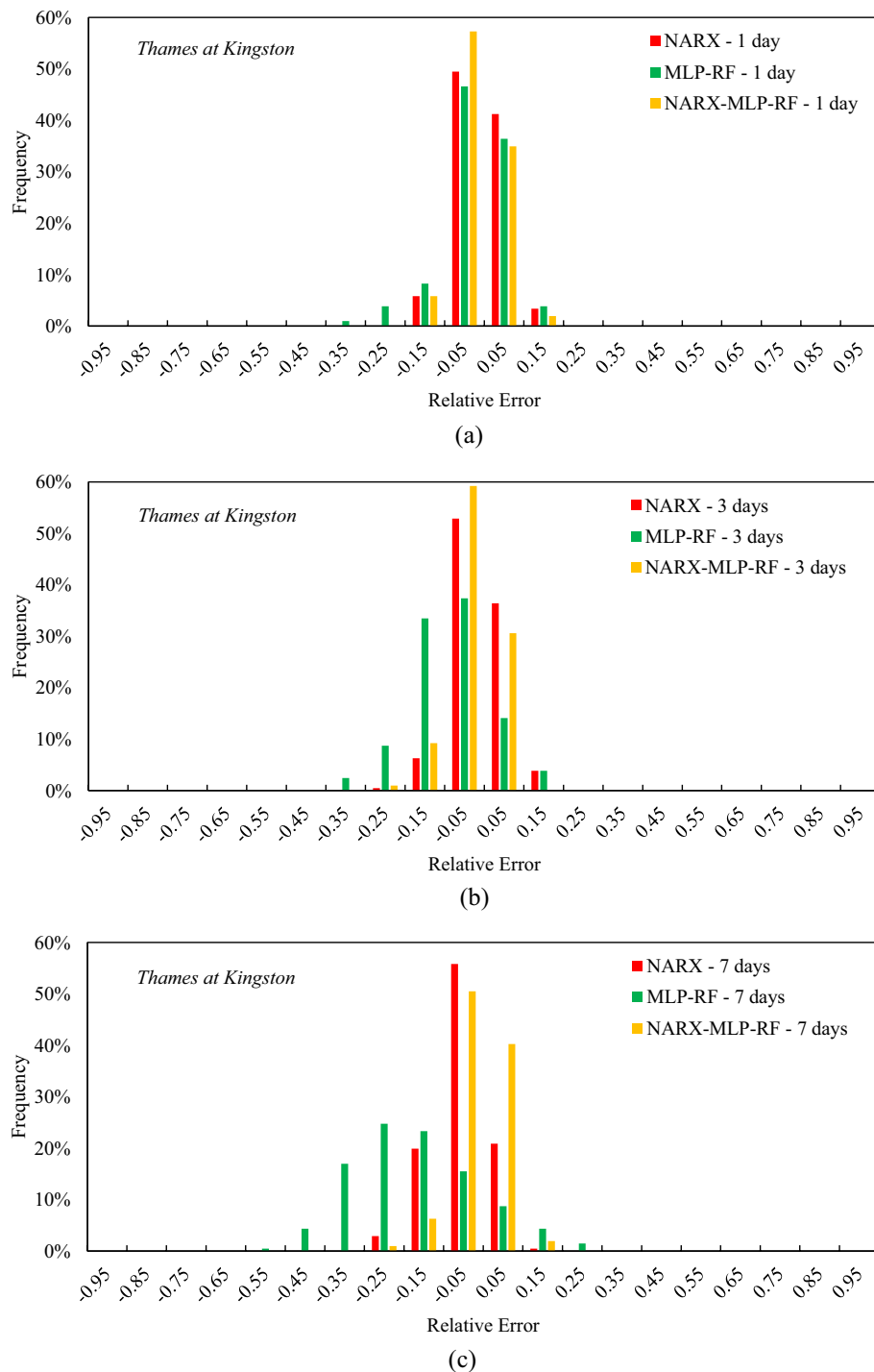


Figure 13. Frequency of the relative error for the tenth decile for Thames at Kingston: $t = 1$ day (a); $t = 3$ days (b); $t = 7$ days (c).

showed a higher frequency of negative relative errors, indicating that underestimates of extreme flows exceed overestimates in terms of frequency. However, the NARX-MLP-RF ensemble still showed a peak frequency of 18% for both the low relative errors between -0.1 and 0 and between 0 and 0.1 .

A lower variance in relative errors was observed for the Thames first-decile flow forecasts in Kingston (Fig. 13), compared to the other two reference rivers. Specifically, for $t = 1$ day, the NARX-MLP-RF ensemble model showed frequencies of 57% and 35% for the lowest relative error between -0.1 and 0 and between 0 and 0.1 , respectively. Furthermore, the relative errors were generally within a narrow range, between -0.2 and 0.2 . MLP-RF showed a slightly worse situation, with a higher frequency of negative relative errors of 8% and 4%, between

−0.2 and −0.1 and between −0.3 and −0.2, respectively. As the forecast horizon increased, the NARX-MLP-RF model still showed an almost symmetric distribution, while both NARX and MLP-RF showed an increase in the frequency of negative relative errors, resulting in a more asymmetric distribution that confirms a greater underestimation of peak flows than the NARX-MLP-RF ensemble model.

Overall, the outcomes observed for streamflow rate prediction preformed on whole time series were in agreement with what observed for the high flows. Actually, while for rivers like the Ribble, with smaller catchments and higher but less homogeneous rainfall throughout the year, relative error ranges were quite wide, for rivers with large catchments and more homogeneous rainfall like the Thames the relative error ranges were narrower, indicating a greater accuracy in the prediction of high flows. However, the hybrid NARX-MLP-RF model proves to be the best, with the NARX and MLP-RF models leading to more asymmetrical distributions even over larger basins.

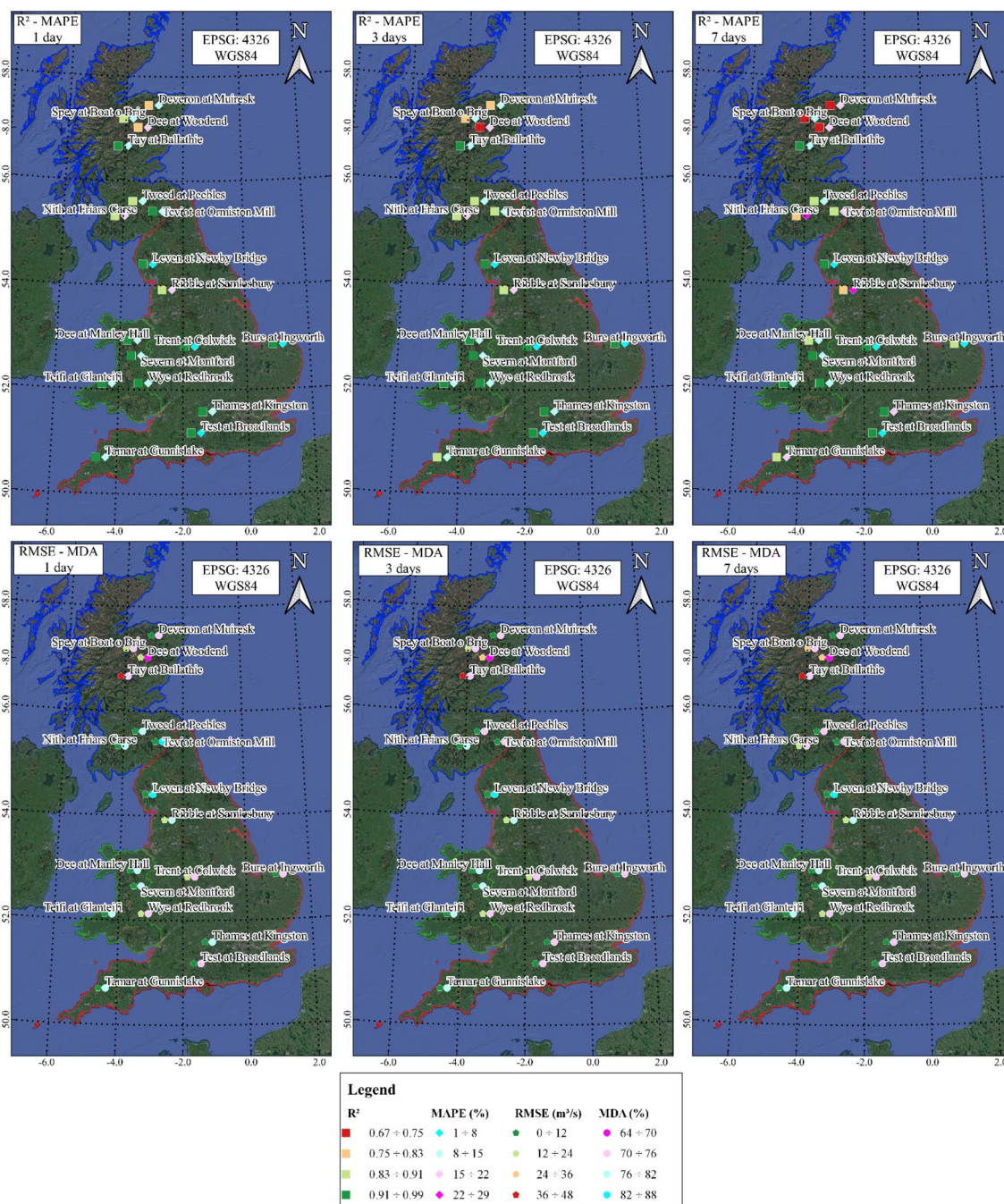


Figure 14. NARX-MLP-RF, testing stage: R²—MAPE (on the top) and RMSE—MDA (on the bottom). Maps created using the Free and Open Source QGIS²⁵.

Streamflow rate predictions for the whole of UK. This section discusses the streamflow forecasts performed with the hybrid NARX-MLP-RF model, with reference to the testing stage, for all investigated rivers. Figure 14 provides a map with the different evaluation metrics, for R^2 -MAPE and RMSE-MDA couples, as the forecast horizon increases. Metrics are also shown in Table 5.

The R^2 coefficient showed values ranging from 0.77 to over 0.99 for the 1-day forecasts. R^2 decreased as the forecast horizon increased, in some cases dropping to values in the order of 0.7 for the 7-day forecast. However, there is a marked territorial difference. For rivers in the south of the UK, an R^2 of over 0.8 was obtained, with peaks as high as 0.95, even for 7-day forecast, while for rivers in Scotland, particularly those in the north-east, lower values of 0.77 and 0.7 were obtained for the 1-day and 7-days ahead predictions, respectively. The MAPE shows a trend in agreement with the R^2 values, with values between 1 and 26%, and increasing with the forecast horizon.

The RMSE values were consistent with the R^2 maps, with lower values for the rivers of England and Wales, ranging from about 4 m³/s to 18 m³/s, and higher values for Scotland. The increase in RMSE as the forecast horizon increased was most pronounced for the northern UK, with RMSE up to about 40 m³/s for 7-days ahead predictions. However, many rivers of England and Wales were characterized by RMSE values between 4 m³/s and 18 m³/s even for 7-days ahead predictions. In addition, MDA values between 64 and 88% were calculated, showing a good ability of the forecasting model to follow the right direction along the streamflow time series. A slight reduction was observed as the forecast horizon increases, with, however values between 64 and 70% observed only for rivers in central and north-east Scotland, where the lowest R^2 values were also obtained.

Overall, the hybrid NARX-MLP-RF model resulted in good predictions for all rivers and forecast horizon. However, the performance of the forecast model is highest for rivers with large basins and a homogeneous distribution of rainfall throughout the year, as observed for several English rivers, while it is lowest for rivers with smaller basins, characterized by less homogeneous rainfall, where peak prediction is more challenging due to the sudden variation in stream flow.

In order to provide an overview of how model performance changes with the forecast horizon, the percentage increase in MAPE, from a 1-day to a 7-day forecast horizon was analysed and reported in Fig. 15.

In particular, the ensemble NARX-MLP-RF model showed the lower MAPE variations for most stations, followed by the NARX model. Both showed MAPE variations of less than 10%. In contrast, MLP-RF showed more marked MAPE variations, with a maximum value of 56% for Tamar at Gunnislake. However, for some stations, MLP-RF also showed MAPE variation of less than 10%. For example, for Test at Broadlands, the MAPE variation was 4.57%. However, for the same station, NARX and NARX-MLP-RF showed lower MAPE variations of 3.60% and 2.90% respectively. It was noted that there is an appreciable correlation between the increase in MAPE just considered and the CV of the flow time series (Fig. 15). The correlation is high for the NARX model ($r = 0.82$) and rather high for the NARX-MLP-RF ensemble model ($r = 0.72$), whereas it is significantly lower for the MLP-RF hybrid model ($r = 0.58$).

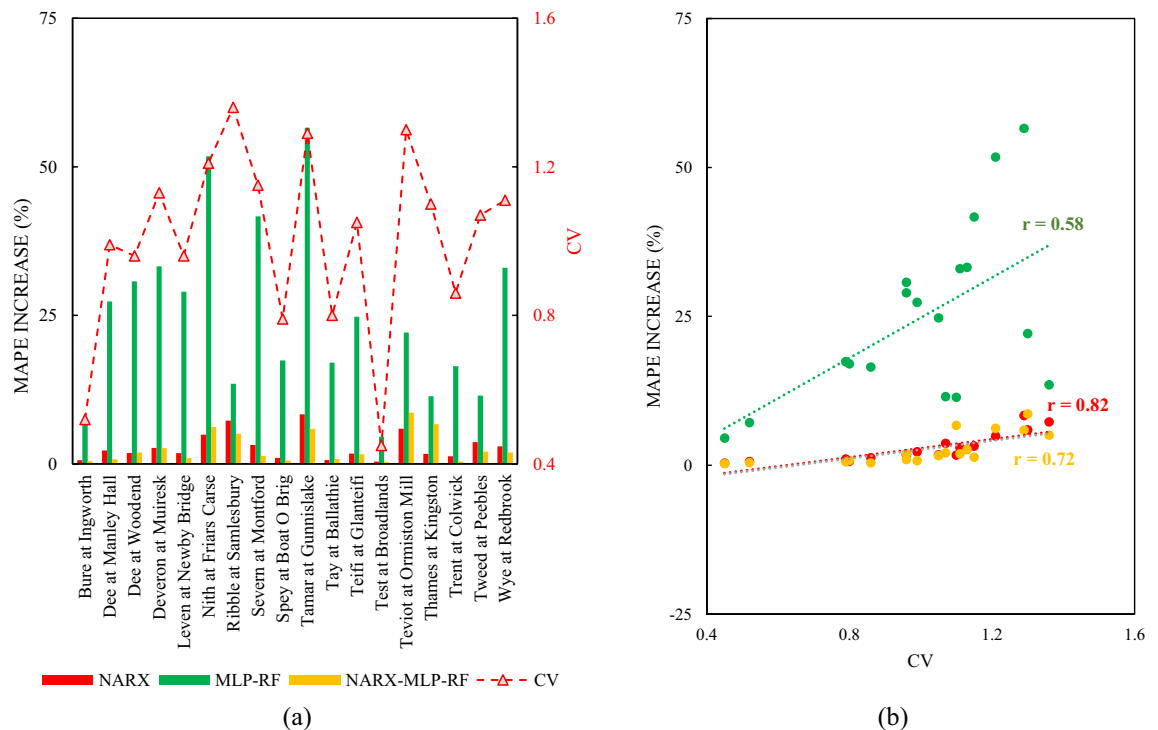


Figure 15. MAPE percentage increase as the forecast horizon increases: histogram for the 18 rivers (a); CV vs MAPE percentage increase (b).

This result demonstrates that while the decrease in accuracy of the forecast models, as the forecast horizon increases, is proportional to the variability of the streamflow during the time series, and this decrease is much less pronounced in the NARX model than in the hybrid RF-MLP one. However, this aspect needs further investigation and specific studies.

Discussion

The extensive study carried out on the streamflow of a large number of rivers in the United Kingdom allows the following to be highlighted:

- The NARX-MLP-RF hybrid model outperformed both the NARX and MLP-RF models for all the investigated rivers and for all forecast horizons. All models resulted in very accurate predictions in the south of UK, while lower performance was observed in the north of UK.
- A reduction in performance was observed as the forecasting horizon increased, but this affected the NARX and MLP-RF models more than the hybrid NARX-MLP-RF model.
- The hybridization of NARX and MLP-RF had a greater impact in improving the predictions obtained for small basins with high and uneven precipitation throughout the year, which make peak forecasting more challenging. Conversely, individual NARX and MLP-RF models led in most cases to satisfactory results, without the need for hybridization, for large basins with a more gradual variation in flow rates.

Regarding the application of hybrid ML models for streamflow forecasting, Li et al.¹³ developed hybrid models for the Yuetan Basin, China, achieving the best results with a PSO-SVR model, which showed a Nash–Sutcliffe Efficiency (which has a mathematical expression almost identical to R^2) of 0.82, lower than the R^2 values obtained for several rivers investigated in this study. The super ensemble model proposed by Tyrallis et al.¹⁵ resulted in large differences, in terms of prediction accuracy, among the large number of investigated rivers in USA, with R^2 values mostly between 0.60 and 0.65. Lee and Ahn³⁹ developed a stacking model based on four ML algorithms: SVM, Gradient Boosting Machine (GBM), Cubist, and Bayesian Regularized Neural Networks (BRNN), for the streamflow rate prediction in South Korea. The authors calculated values of NSE up to 0.48, also showing a performance reduction as the forecast horizon increased, as observed in the present study. Kilinc and Yurtsever⁴⁰ also developed a hybrid DL model Based on Grey Wolf algorithm (GWO) and GRU for the daily streamflow forecasting in two stations located in the Seyhan basin, Turkey. The authors showed the advantages of the hybridization based on DL algorithms, achieving accurate predictions with R^2 values up to 0.98. The results obtained by Kilinc and Yurtsever⁴⁰ are in line with the prediction obtained for several rivers investigated in the present study for $t = 1$ day. However, they did not perform an analysis with increasing time horizon, as made in the present study. Granata et al.²¹, who proposed a comparison between Bi-LSTM and a stacked MLP-RF model, obtained very accurate predictions, also for the UK Trent River investigated in this study. However, they also showed a reduction in prediction accuracy as the forecast horizon increased, already for the 3-days forecast.

A comparison was also made with literature studies investigating the impact of climatic factors and catchment characteristics on the accuracy of river discharge forecasts. Xu et al.⁴¹ investigated the spatial and temporal scale effects on the predictive performance of the monthly streamflow prediction, based on a hybrid DL model based on the CNN and GRU algorithms applied to many watersheds around globe. The authors showed how the hybrid DL model performs better on large drainage areas, in agreement with the present study. Moreover, the predictive performance tends to get better also with the extension of a training period for the model, confirming how long time series can lead to more accurate predictions. Harrigan et al.⁴² evaluating the Ensemble Streamflow Prediction (ESP) method for 314 catchments in the UK, exploring the relationship between basins characteristics and ESP skill. The ESP method allows factors such as precipitation, potential evapotranspiration, temperature, soil moisture, groundwater and snow for each basin to be included in the modelling. The authors showed how the performance of the ESP model decayed exponentially with increasing forecast horizon, but large catchments decayed at a slower rate. In addition, better performances were observed in the south and east of the UK, where large and slower responding catchments are mainly located. Conversely, lower performances were observed for the highly responsive catchments in the north and west. These outcomes are in agreement with the present study. We showed that for large basins, such as for the Thames River in southern England, the models tested led to accurate predictions for both ordinary and high flows, whereas for smaller basins, such as for the Ribble River in Northern England, forecasts were less accurate and decayed in accuracy at a higher rate compared to larger basins as the forecast horizon increased, particularly for the NARX and MLP-RF models.

Overall, although the methodology has been tested on a significant number of rivers, UK weather and climate conditions have different features in comparison with warmer climates. In the future it will be interesting to test the methodology in semi-arid and Mediterranean areas, where the seasonal pattern of rainfall is more pronounced compared to UK. From this perspective, different ML or deep-learning algorithms could be included, together with further exogenous inputs, in the forecast procedure in order to improve the reliability of the streamflow rate forecasting. This could lead to overcoming the current limitations related to climate and streamflow regimes on the one hand and the forecasting horizon on the other, moving from the current short-term to the medium-term scenario.

Conclusion

A novel streamflow prediction model, for forecast horizons of up to seven days, was developed in this research and applied to a regional study that considered 18 rivers throughout the UK. The proposed model was obtained by stacking the NARX, RF, and MLP algorithms and used a BO procedure for tuning the hyperparameters.

Daily precipitations were considered as the only exogenous input variable. The NARX-MLP-RF ensemble model showed very good forecasting capabilities and outperformed both NARX and MLP-RF models, for all rivers and forecast horizons. NARX-MLP-RF showed a lower reduction of accuracy as the forecasting horizon increased, for both regular and extreme streamflow, compared to the NARX and MLP-RF models. In this regard, for both the NARX-MLP-RF and NARX models, a significant correlation was found between the increase in MAPE corresponding to the increase in the forecast horizon from 1 to 7 days and the CV of the flow time series.

In addition, NARX-MLP-RF has proven to be particularly suitable for providing accurate forecasts for rivers with small catchment areas with highly variable rainfall and streamflow rate distributions over time., for which the forecasting of the often-abrupt peaks is a challenging task. In particular, more accurate forecast values were generally obtained for rivers in Wales and southern England.

Overall, the accurate predictions made with the NARX-MLP-RF model make it a powerful tool for managing the risks associated with possible extreme flows involving frequent floods, and also for short-term water management decision-making.

Data availability

Data from the National River Flow Archive, which is the primary archive of daily and peak river flows for the United Kingdom, were used in the creation of this manuscript. Data are available at the following website: <https://nrfa.ceh.ac.uk>. The elaborations were carried out mostly with the following software: MATLAB (<https://mathworks.com>), Microsoft Excel (<https://www.microsoft.com/en-ww/microsoft-365/excel>), and QGIS (<https://qgis.org/en/site>).

Received: 24 January 2023; Accepted: 27 April 2023

Published online: 29 April 2023

References

- Kuriqi, A. *et al.* Seasonality shift and streamflow flow variability trends in central India. *Acta Geophys.* **68**, 1461–1475. <https://doi.org/10.1007/s11600-020-00475-4> (2020).
- Arnold, J. G., Srinivasan, R., Muttiah, R. S. & Williams, J. R. Large area hydrologic modeling and assessment. Part I. Model development. *J. Am. Water Resour. Assoc.* **34**, 73–89. <https://doi.org/10.1111/j.1752-1688.1998.tb05961.x> (1998).
- Shen, C. & Phanikumar, M. S. A process-based, distributed hydrologic model based on a large-scale method for surface–subsurface coupling. *Adv. Water Resour.* **33**(12), 1524–1541. <https://doi.org/10.1016/j.advwatres.2010.09.002> (2010).
- Kostić, S., Stojković, M., Prohaska, S. & Vasović, N. Modeling of river flow rate as a function of rainfall and temperature using response surface methodology based on historical time series. *J. Hydroinf.* **18**(4), 651–665. <https://doi.org/10.2166/hydro.2016.153> (2016).
- Granata, F. & Di Nunno, F. Forecasting evapotranspiration in different climates using ensembles of recurrent neural networks. *Agric. Water Manag.* <https://doi.org/10.1016/j.agwat.2021.107040> (2021).
- Granata, F. & Di Nunno, F. Artificial Intelligence models for prediction of the tide level in Venice. *Stoch. Env. Res. Risk Assess.* <https://doi.org/10.1007/s00477-021-02018-9> (2021).
- Pham, Q. B. *et al.* Groundwater level prediction using machine learning algorithms in a drought-prone area. *Neural Comput. Appl.* **34**(13), 10751–10773. <https://doi.org/10.1007/s00521-022-07009-7> (2022).
- Kiş, Ö. River flow forecasting and estimation using different artificial neural network techniques. *Hydrol. Res.* **39**(1), 27–40. <https://doi.org/10.2166/nh.2008.026> (2008).
- Galavi, H., Mirzaei, M., Shui, L. T. & Valizadeh, N. Klang river level forecasting using ARIMA and ANFIS models. *J. Am. Water Works Assoc.* **105**(9), E496–E506. <https://doi.org/10.5942/jawwa.2013.105.0106> (2013).
- Yaseen, Z. M., El-Shafie, A., Jaafar, O., Afan, H. A. & Sayl, K. N. Artificial intelligence based models for stream-flow forecasting: 2000–2015. *J. Hydrol.* **530**, 829–844. <https://doi.org/10.1016/j.jhydrol.2015.10.038> (2015).
- Khan, U. T. & Valeo, C. Short-term peak flow rate prediction and flood risk assessment using fuzzy linear regression. *J. Environ. Inf.* **28**(2), 71–89. <https://doi.org/10.3808/jei.201600345> (2016).
- Elbeltagi, A., Di Nunno, F., Kushwaha, N. L., de Marinis, G. & Granata, F. River flow rate prediction in the Des Moines watershed (Iowa, USA): A machine learning approach. *Stoch. Env. Res. Risk Assess.* <https://doi.org/10.1007/s00477-022-02228-9> (2022).
- Li, X., Sha, J., Li, Y. & Wang, Z. L. Comparison of hybrid models for daily streamflow prediction in a forested basin. *J. Hydroinf.* **20**, 191–205. <https://doi.org/10.2166/hydro.2017.189> (2018).
- Saraiva, S. V., de Oliveira Carvalho, F., Santos, C. A. G., Barreto, L. C. & Freire, P. K. D. M. M. Daily streamflow forecasting in Sobradinho Reservoir using machine learning models coupled with wavelet transform and bootstrapping. *Appl. Soft Comput.* **102**, 107081–107116. <https://doi.org/10.1016/j.asoc.2021.107081> (2021).
- Tyralis, H., Papacharalampous, G. & Langousis, A. Super ensemble learning for daily streamflow forecasting: Large-scale demonstration and comparison with multiple machine learning algorithms. *Neural Comput. Appl.* **33**(8), 3053–3068. <https://doi.org/10.1007/s00521-020-05172-3> (2021).
- Kumar, M. *et al.* Estimation of daily stage–discharge relationship by using data-driven techniques of a perennial river India. *Sustainability* **12**(19), 7877. <https://doi.org/10.3390/su12197877> (2020).
- Kumar, M., Kumar, P., Kumar, A., Elbeltagi, A. & Kuriqi, A. Modeling stage–discharge–sediment using support vector machine and artificial neural network coupled with wavelet transform. *Appl. Water Sci.* **12**, 87. <https://doi.org/10.1007/s13201-022-01621-7> (2022).
- Fu, M. *et al.* Deep learning data-intelligence model based on adjusted forecasting window scale: Application in daily streamflow simulation. *IEEE Access* **8**, 32632–32651. <https://doi.org/10.1109/ACCESS.2020.2974406> (2020).
- Le, X. H., Nguyen, D. H., Jung, S., Yeon, M. & Lee, G. Comparison of deep learning techniques for river streamflow forecasting. *IEEE Access* **9**, 71805–71820. <https://doi.org/10.1109/ACCESS.2021.3077703> (2021).
- Ahmed, A. M. *et al.* Deep learning hybrid model with Boruta–Random forest optimiser algorithm for streamflow forecasting with climate mode indices, rainfall, and periodicity. *J. Hydrol.* **599**, 126350. <https://doi.org/10.1016/j.jhydrol.2021.126350> (2021).
- Granata, F., Di Nunno, F. & de Marinis, G. Stacked machine learning algorithms and bidirectional long short-term memory networks for multi-step ahead streamflow forecasting: a comparative study. *J. Hydrol.* **613**(1–4), 128431. <https://doi.org/10.1016/j.jhydrol.2022.128431> (2022).
- Wegayehu, E. B. & Muluneh, F. B. Short-term daily univariate streamflow forecasting using deep learning models. *Adv. Meteorol.* <https://doi.org/10.1155/2022/1860460> (2022).

23. Hassan Ibrahim, K. S. M., Huang, Y. F., Ahmed, A. N., Koo, C. H. & El-Shafie, A. A review of the hybrid artificial intelligence and optimization modelling of hydrological streamflow forecasting. *Alex. Eng. J.* **61**(1), 279–303. <https://doi.org/10.1016/j.aej.2021.04.100> (2022).
24. Di Nunno, F., Granata, F., Gargano, R. & de Marinis, G. Prediction of spring flows using nonlinear autoregressive exogenous (NARX) neural network models. *Environ. Monit. Assess.* <https://doi.org/10.1007/s10661-021-09135-6> (2021).
25. QGIS Development Team. QGIS geographic information system. Version 3.28.5. Open Source Geospatial Foundation Project. <http://qgis.osgeo.org> (2023).
26. Li, S., Zhang, L., Du, Y., Zhuang, Y. & Yan, C. Anthropogenic impacts on streamflow-compensated climate change effect in the Hanjiang River basin China. *J. Hydrol. Eng.* **25**(1), 04019058. [https://doi.org/10.1061/\(ASCE\)HE.1943-5584.0001876](https://doi.org/10.1061/(ASCE)HE.1943-5584.0001876) (2020).
27. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. Royal Stat. Soc. Series B (Stat. Methodol.)* **67**(2), 301–320. <https://doi.org/10.1111/j.1467-9868.2005.00503.x> (2005).
28. Hastie, T., Tibshirani, R. and Friedman J. The elements of statistical learning: Data mining, inference, prediction. Springer Series in Statistics, https://doi.org/10.1007/978-0-387-84858-7_8 (2009).
29. Wu, J. *et al.* Hyperparameter optimization for machine learning models based on Bayesian optimization. *J. Electron. Sci. Technol.* **17**(1), 26–40. <https://doi.org/10.11989/JEST.1674-862X.80904120> (2019).
30. Snoek, J., Larochelle, H. and Adams, R. P. Practical bayesian optimization of machine learning algorithms. In: *Advances in Neural Information Processing Systems*, p. 25, (2012).
31. Di Nunno, F., Race, M. & Granata, F. A nonlinear autoregressive exogenous (NARX) model to predict nitrate concentration in rivers. *Environ. Sci. Pollut. Res.* **29**, 40623–40642. <https://doi.org/10.1007/s11356-021-18221-8> (2022).
32. Boussaada, Z., Curea, O., Remaci, A., Camblong, H. & Mrabet Bellaaj, N. A nonlinear autoregressive exogenous (NARX) neural network model for the prediction of the daily direct solar radiation. *Energies* **11**(3), 620. <https://doi.org/10.3390/en11030620> (2018).
33. MacKay, D. J. C. Bayesian Interpolation. *Neural Comput.* **4**, 415–447. <https://doi.org/10.1162/neco.1992.4.3.415> (1992).
34. Di Nunno, F. & Granata, F. Groundwater level prediction in Apulia region (Southern Italy) using NARX neural network. *Environ. Res.* **190**, 110062 (2020).
35. Di Nunno, F., Granata, F., Gargano, R. & de Marinis, G. Forecasting of extreme storm tide events using NARX neural network-based models. *Atmosphere* **12**(4), 512. <https://doi.org/10.3390/atmos12040512> (2021).
36. Rosenblatt, F. Principles of neurodynamics. Perceptrons and the theory of brain mechanisms. Cornell Aeronautical Lab Inc Buffalo NY, (1961).
37. Murtagh, F. Multilayer perceptrons for classification and regression. *Neurocomputing* **2**(5–6), 183–197 (1991).
38. Breiman, L. Random forests. *Mach. Learn.* **45**(1), 5–32 (2001).
39. Lee, D. G. & Ahn, K. H. A stacking ensemble model for hydrological postprocessing to improve streamflow forecasts at medium-range timescales over South Korea. *J. Hydrol.* **600**, 126681. <https://doi.org/10.1016/j.jhydrol.2021.126681> (2021).
40. Kilinc, H. C. & Yurtsever, A. Short-term streamflow forecasting using hybrid deep learning model based on grey wolf algorithm for hydrological time series. *Sustainability* **14**(6), 3352. <https://doi.org/10.3390/su14063352> (2022).
41. Xu, W., Chen, J. & Zhang, X. J. Scale effects of the monthly streamflow prediction using a state-of-the-art deep learning model. *Water Resour. Manage* **36**, 3609–3625. <https://doi.org/10.1007/s11269-022-03216-y> (2022).
42. Harrigan, S., Prudhomme, C., Parry, S., Smith, K. & Tanguy, M. Benchmarking ensemble streamflow prediction skill in the UK. *Hydrol. Earth Syst. Sci.* **22**, 2023–2039. <https://doi.org/10.5194/hess-22-2023-2018> (2018).

Acknowledgements

F.D.N. contributed to the data collection, to the development of forecast models and to the manuscript preparation; F.G. contributed to the conceptualization of the study, to the development of forecast models and to the manuscript preparation; G.d.M. contributed to the administration of the research project and to the manuscript preparation.

Author contributions

F.D.N. contributed to the data collection, to the development of forecast models and to the manuscript preparation; F.G. contributed to the conceptualization of the study, to the development of forecast models and to the manuscript preparation; G.d.M. contributed to the administration of the research project and to the manuscript preparation.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to F.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023