Short-Term Streamflow Forecasting for Paraíba do Sul River Using Deep Learning

Luciana Conceição Dias Campos, Leonardo Goliatt da Fonseca(✉), Tales Lima Fonseca, Gabriel Dias de Abreu, Letícia Florentino Pires, and Yulia Gorodetskaya

Federal University of Juiz de Fora, Juiz de Fora, MG, Brazil
lcdcampos@gmail.com, goliatt@gmail.com, taleslimaf@gmail.com, GABRIEL199716@gmail.com, lepirescomp@gmail.com, yu.gorodetskaya@gmail.com

Abstract. Water resources are essential for sustainable economic and social development, as well as be a vital element for the conservation of ecosystems and the life of all beings on our planet. On the other hand, natural and anthropic disasters from floods and droughts may occur. The modeling of hydrological historical series has extensively been studied in the literature for important applications involving the water resources' planning and management. There are several temporal series prediction's techniques in the literature. Some of them are characterized as classical linear methods whose adjusts for multivariate or multi-input prediction problems can be difficult. On the other hand, artificial neural networks can learn complex nonlinear relationships from time series, and the deep learning model LSTM is considered the most successful type of recurrent neural network capable of directly supporting multivariate prediction problems. This work presents a comparison between two forecasting's models of time series: ARIMA, a classical linear model, and an LSTM neural network, a nonlinear model. As a case study, we used the time series of four measurings' substations of one of the very important Brazilian rivers - the Paraíba do Sul river. These time series are difficult to predict since their history series has flaws and high oscillation in the data. The LSTM, which is a robust model, performs better in analyzing the behavior of this type of time series.

Keywords: Deep learning · Long short-term memory · Time series

1 Introduction

The effective management of available water resources in a river basin requires several aspects, including proper models to be as accurate as possible in predicting future outflows. The economic development and life of the vast majority of people rely on these water resources, which increases the need for improvement in their administration tools. The modeling of hydrological historical series

P. Moura Oliveira et al. (Eds.): EPIA 2019, LNAI 11804, pp. 507–518, 2019.
https://doi.org/10.1007/978-3-030-30241-2_43

has been extensively studied in the literature for important applications such as drought management, water quality policies, flood forecasting, electric generation, environmental management, water services, and more efficient land use in human, industrial and agricultural supply [3,6,12,17].

In recent decades, artificial neural networks have been a valuable tool in many areas of research. The uses of these techniques have also gained ground in the area of water resources, where, in most times, its use in flow prediction presented results compatible or superior to traditional techniques [4,15,21,26,29,33].

Artificial neural networks (RNA) contain adaptive weights along paths between neurons, which can be adapted by a learning algorithm, whose learning occurs through data that are observed until the prediction of the model has a satisfactory result [35].

These techniques are capable of representing complex and non-linear relationships, as well as investigating phenomena from multiple data sources and transform forecast simulations for a data-based practice.

A breakthrough in the artificial neural network field is Deep Learning. Deep-learning networks are distinguished from the more commonplace single-hidden-layer neural networks by their depth; that is, the number of node layers through which data must pass in a multistep process of data's learning. A special kind of recurrent neural network (RNN) architecture used in the field of Deep Learning is Long Short Term Memory networks – called "LSTMs". They are capable of learning long-term dependencies and they work tremendously well on a large variety of problems, like as Natural language processing [24], speech recognition [11] and extreme events requesting [22].

LSTM networks are well-suited to classifying, processing and making predictions based on time series data since there can be lags of unknown duration between important events in a time series. LSTMs were developed to deal with the exploding and vanishing gradient problems that can be encountered when training traditional RNNs, which motivated this work [20,30,36].

This paper aims to assess the viability of using LSTM in forecasting short-term streamflows for time series. To assess the performance of LSTM, we have used historical data from the Paraíba do Sul River (PSR), Brazil. The PSR basin has as main economic activities the industrial and agricultural sectors, and it is characterized by conflicts of multiple uses of its water resources [28,34]. In the last few years, the research on the PSR has been gained increasing attention in several areas. Some efforts include the research on degradation of the aquatic system due to human activities [16], dynamics of sediment transportation [23,32], geological studies [5] and drought identification and characterisation [31]. The prediction of the natural flow of PSR is one of the most important factors for analyses involving the management of this basin. In this way, accurate forecasting tools are essential for a robust and reliable decision-making process.

Historical data from PSR basin have high oscillating and missing data. The success of LSTM in dealing with these data comes from its capacity to capture long-range dependencies over time, which is acquired by the structure of its cell.

The cell is composed of an update gate, output gate and forget gate. These three gates regulate the flow of information into and out of the cell [14].

The secondary aim of the study accomplishes a comparison between ARIMA model [27], a well established statistical approach to streamflow forecasting and the LSTM model, a data-driven non-linear model.

The results of this work show the robust performance of the LSTM model compared to a classic ARIMA model, ensuring that LSTM adheres well to the prediction of the PSR river flow problem.

The rest of this paper is organized as follows: Sect. 2 presents the fundamental concepts around the subject which give context to other sections. Section 3 presents the executed experiment and the results. Section 4 shows the final results and supplies the reader with our contributions to the literature, promoting discussion around the research method with future suggestions.

2 Material and Methods

2.1 Dataset

The dataset used in this paper was provided by the Brazilian National Water Agency (ANA) and compounded by four daily measurement stations located on Paraíba do Sul's river basin. These stations are referenced as 58218000 (UHE FUNIL MONTANTE 2), 58235100 (QUELUZ), 58880001 (SÃO FIDELIS), 58974000 (CAMPOS - PONTE MUNICIPAL) and keeps observations from 1920 until 2016 [2].

Figure 1 shows the location of the Paraíba do Sul river basin and the four cited fluviometric/rainfall stations.

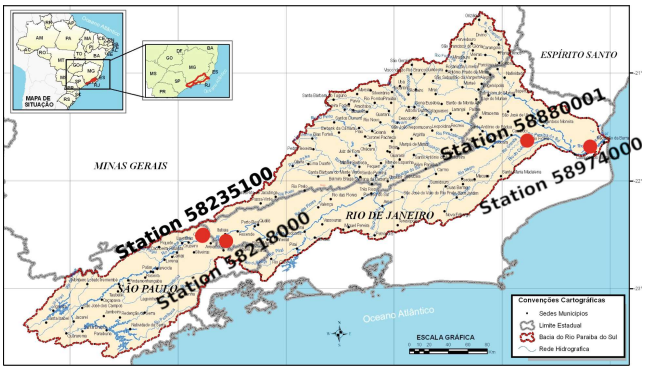

Fig. 1. Location of the Paraíba do Sul river basin and the four fluviometric/rainfall stations. Adapted from [1].

Figure 2 presents time series of the four cited measure stations. It's possible to see, at the raw data, that these series have inconsistencies in their values,

containing missing values with long time gaps without registration. In this work, the missing values are treated with the imputation of median values to keep every observation of the series and them improving LSTM performance, once it fits the data.

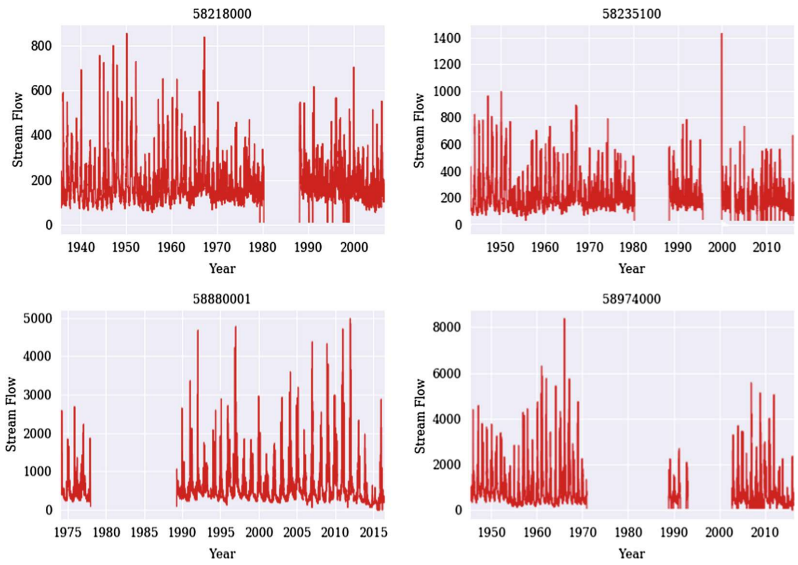

Fig. 2. Streamflow raw data of each measure station, supplied from the Brazilian Agency ANA [2]

2.2 Long Short Term Memory (LSTM)

LSTM is a special type of RNN which are networks with structures of the internal self-looped cell, that allows them to capture data's dynamic temporal behavior. In cases of architectures such as LSTMs, they can capture long time sequences, hence they have been used for time series with success.

LSTM as in Fig. 3 address this problem with an input gate, that specifies the information that will be stored to the cell state, an output gate that specifies which information from cell state will flow as output and a forget gate that decides which information will be removed from cell state. Therefore, these mechanisms from LSTM allows information easily flow through the cells unchanged, providing better learning of long-term dependencies [9].

2.3 Autoregressive Integrated Moving Average (ARIMA)

The autoregressive integrated moving average (ARIMA) models are the most general class of models for forecasting a time series that can to become "stationary" by differencing (when necessary), perhaps in conjunction with nonlinear

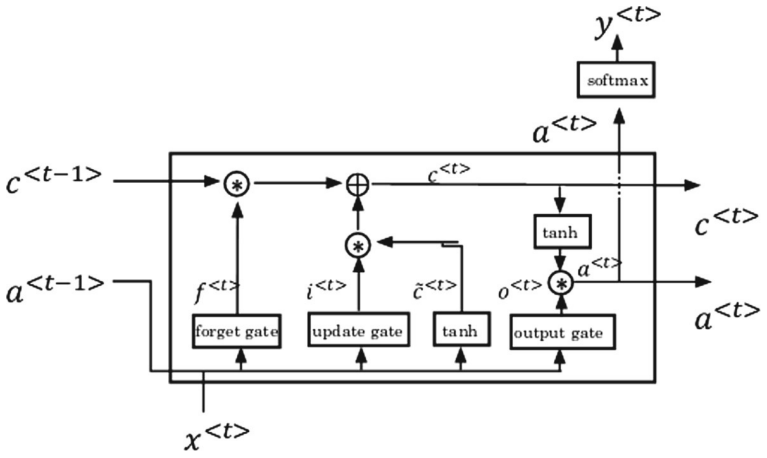

Fig. 3. LSTM cell [25].

transformations such as logging or deflating (when necessary). The process of fitting an ARIMA model is sometimes referred to as the Box-Jenkins method [8].

Lags of the differenced series appearing in the forecasting equation are called "auto-regressive" terms (AR), lags of the forecast errors are called "moving average" terms (MA), and a time series which needs to be differenced to be made stationary is said to be an "integrated" version of a stationary series (I). Each of these components is explicitly specified in the model as a parameter, characterized by 3 terms: p is the order of the AR term, q is the order of the MA term and d is the number of differences required to make the time series stationary.

Given a time series of data X_t where t is an integer index and the X_t are real numbers, the notation ARIMA (p, d, q) model:

1. First, let x denote the d_{th} difference of X, which means:
   - If $d = 0$: $x_t = X_t$.
   - If $d = 1$: $x_t = X_t - X_{t-1}$
   - If $d = 2$: $x_t = (X_t - X_{t-1}) - (X_{t-1} - X_{t-2}) = X_t - 2X_{t-1} + X_{t-2}$
2. In terms of x, the general forecasting equation is present at the Eq. 1:

$$\hat{x}_t = \mu + \phi_1 x_{t-1} + \ldots + \phi_p x_{t-p} - \theta_1 \epsilon_{t-1} - \ldots - \theta_q \epsilon_{t-q} \tag{1}$$

where $\hat{x}_t$ is the forecast of the time series at time t, $\phi_1 \ldots \phi_p$ and $\theta_1 \ldots \theta_q$ are the parameters of the model, and $\epsilon_{t-1} \ldots \epsilon_{t-q}$ is the residual error series and they are white noises, that is, the residuals themselves are independent and identically distributed (i.i.d.).

2.4 Time Series Cross-validation (TSCV)

Cross-validation is a statistical sampling technique to evaluate the generalization of a model from a data set [19]. K-Fold [13] is one of the most used meth-

ods among the cross-validation techniques, it eliminates dependency on subsets of validation that compromises most general model selection preventing overfitting.

In a Time series scope, a variation of K-Fold technique that is called Time Series Split (TSS) permits this same generalization gains without shuffling observations.

TSCV as illustrated by Fig. 4 splits all train data in n consecutive samples with train followed by validation, on each split, it maintains already trained and validation subsets as new train data and aggregates another split dividing in train and validation subsets. Validation subsets are the same sized and train keeps growing at each step, in the end, it selects the best model measured by validation set.

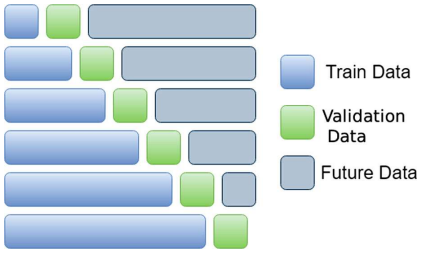

Fig. 4. Time series split example with $k = 6$.

3 Streamflow Estimation Model with LSTM

To carry out the predictions, we selected periods with 14 days in the historical flow series. At every step in the moving window predicting the seventh day ahead.

The predicted flow was considered as a function of finite sets of antecedent flow observations at the stations. The predictive model has the following form:

$$Q_{t+7} = F(Q_t, Q_{t-1}, \cdots, Q_{t-13}) \tag{2}$$

where Q_{t+j} is the streamflow at day $t + j$ and F is an estimation function.

Figure 5 depicts the framework of the proposed approach. The missing data are imputed by the median of the time series, 14-sized rolling windows are generated from data, the last 30% of data is reserved to test set and rest is to train and validate models. The TSCV is the cross-validation method to select the best model and prevent overfitting, this process culminates in the final MAPE evaluation in the test set. This process is performed 30 times resulting in averaged MAPE.

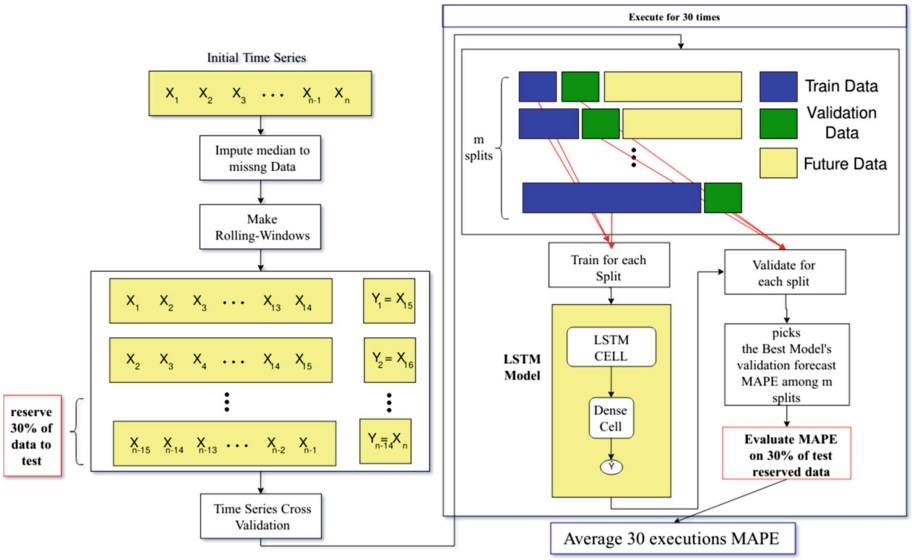

Fig. 5. Streamflow estimation model

3.1 Description of Experiments

ARIMA hyper-parameters as shown in Table 1 was chosen by a grid search with the best in-sample result determining the predictor models configuration as in [27].

Table 1. Hyper-parameters for the ARIMA model at each station.

Station	P	D	Q
58880001	3	0	1
58974000	2	1	0
58218000	2	0	3
58235100	1	0	0

LSTM hyperparameters were set as follows: kernel was initialized in the hidden layer using Xavier algorithm in order to achieve fast convergence [10], mini-batch size equals to 256 as suggested in the literature, and 200 neurons in hidden-layer.

LSTM layer needs two non-linear activation functions that were set to $hard-sigmoid$ and hyperbolic tangent ($tanh$) as defined in [14] to avoid gradient vanishing problem. The last Layer with one neuron to forecast was composed of a linear activation.

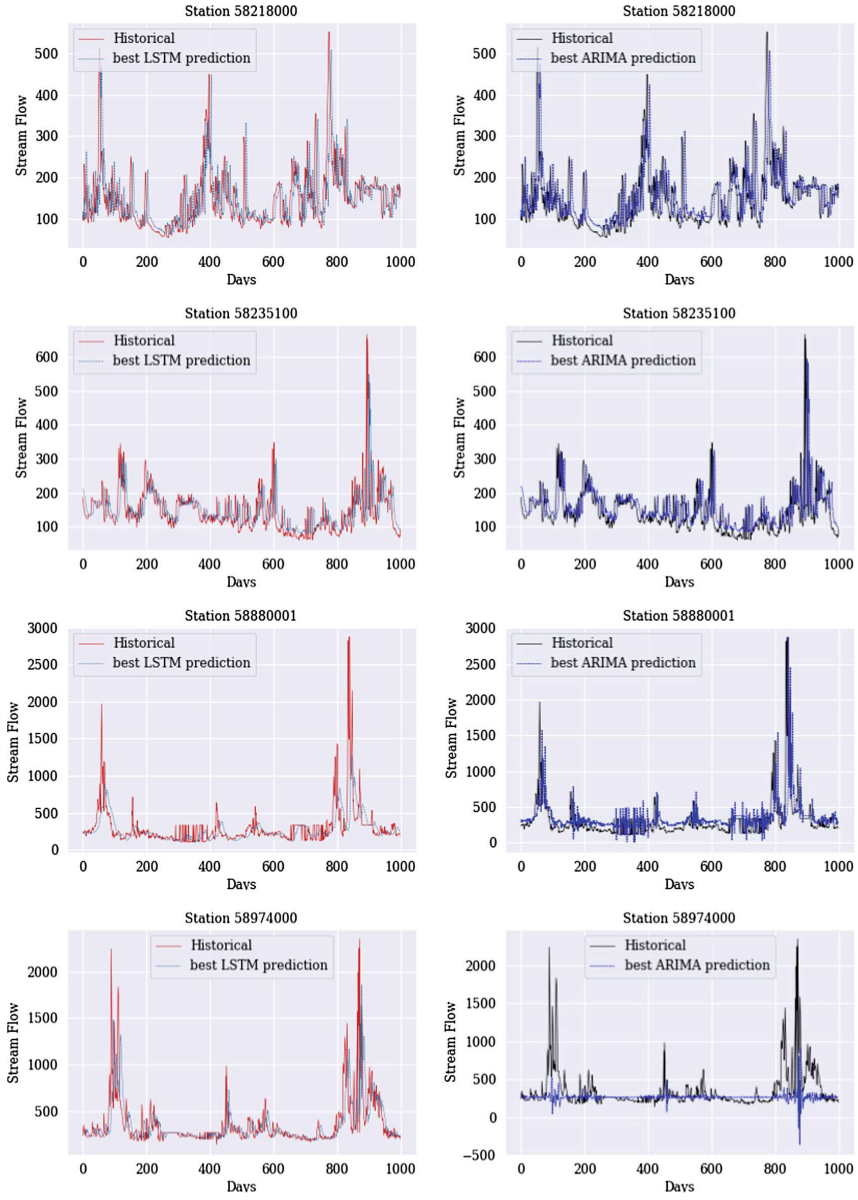

Fig. 6. Last 1000 days of the test set with the best LSTM forecast and historical values at left, and the best ARIMA forecast and historical values at right.

This model was optimized by Adam algorithm [18], time series cross-validation and early-stopping were applied to prevent overfitting [7].

Table 2. Comparison of models MAPEs.

Station	ARIMA	LSTM	Best LSTM
58880001	32.78	27.73 ± 1.01	**25.83**
58974000	26.43	21.29 ± 0.76	**19.74**
58218000	18.53	18.20 ± 0.24	**17.60**
58235100	17.61	15.26 ± 0.32	**14.77**

The main objective of this study was to evaluate the performance of the LSTM in the forecast of the flow of the time series of four stations of measurement of the Paraíba do Sul river. These series, represented in Fig. 2, record strong oscillations, besides present long periods with missing data.

Forecasting in these conditions is a tough mission but we found evidence that LSTM effectively was capable to learn the adversity in this series and predict reasonably well as indicated by Table 2 and Fig. 6.

Figure 6 shows an important problem that this study confronted, the LSTM forecast has some delay when compared to the historical time series. This problem occurs because LSTM is limited by past days, in this sense, we suggest increasing information supply with exogenous variables for future researches.

A limitation of the machine learning approach employed here is that the estimated river flows are reliable only under conditions similar to those that such models have historically experienced. The use of these models to generate predictions in conditions that exceed historical variability may introduce considerable uncertainty into their flow forecasts.

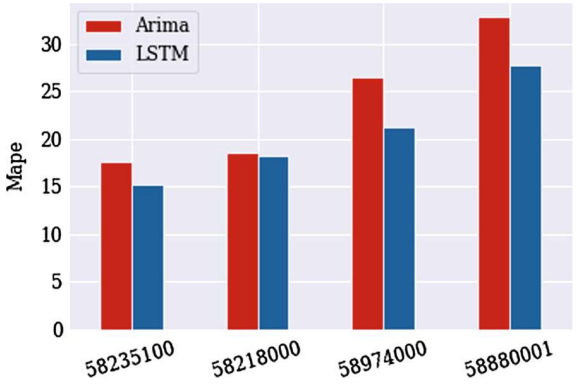

Fig. 7. Comparison of models *MAPE* for each station

In sum of that, Arima as a well-known statistical model to forecast flow performed considerably poor in comparison to LSTM in three of the four stations as saw in Fig. 7 with LSTM out-performing in every Station which corroborates the statement that artificial intelligence algorithms can capture the noise complexity, non-stationarity, dynamism and non-linearity in the data as reported in [35].

4 Conclusions

This work proposes to make the short-term forecast for the time series of the Paraíba River the South through deep learning, more precisely, a model using LSTM. The results of the experiments show that LSTM is a viable approach to prediction time series with high oscillation in data and long periods of missing data. Thus, the prediction results of the model using LSTM surpassed the ARIMA model, a statistical model widely used in forecasting time series of flow.

References

1. Governo do Brasil. https://www.brasil.gov.br/noticias/meio-ambiente. Accessed 27 Mar 2019
2. National water agency. https://www.ana.gov.br/. Accessed 21 Jun 2019
3. Abudu, S., Cui, C.I., King, J.P., Abudukadeer, K.: Comparison of performance of statistical models in forecasting monthly streamflow of Kizil river, China. Water Sci. Eng. **3**(3), 269–281 (2010)
4. Asadi, S., Shahrabi, J., Abbaszadeh, P., Tabanmehr, S.: A new hybrid artificial neural networks for rainfall-runoff process modeling. Neurocomputing **121**, 470–480 (2013)
5. Carelli, T.G., Plantz, J.B., Borghi, L.: Facies and paleoenvironments in paraíba do sul deltaic complex area, north of Rio de Janeiro state. Brazil. J. South American Earth Sci. **86**, 431–446 (2018)
6. Carlisle, D.M., Falcone, J., Wolock, D.M., Meador, M.R., Norris, R.H.: Predicting the natural flow regime: models for assessing hydrological alteration in streams. River Res. Appl. **26**(2), 118–136 (2010)
7. Caruana, R., Lawrence, S., Giles, C.L.: Overfitting in neural nets: backpropagation, conjugate gradient, and early stopping. In: Advances in Neural Information Processing Systems, pp. 402–408 (2001)
8. George, E.P., Box, G.M.J.: Time Series Analysis: Forecasting and Control. Holden-Day Series in time series analysis and digital processing. Holden-Day, San Francisco (1976)
9. Gers, F., Schmidhuber, J., Cummins, F.: Learning to forget: continuous prediction with LSTM. Technical report, Technical Report IDSIA-01-99 (2000)
10. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 249–256 (2010)
11. Graves, A., Jaitly, N., Mohamed, A.R.: Hybrid speech recognition with deep bidirectional LSTM. In: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 273–278. IEEE (2013)

12. Guimarãlise da previsibilidade de cheias na bacia do rio uruguai através do modelo mgb-iph (2018)
13. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning - Data Mining, Inference, and Prediction, 2nd edn. Springer, New York (2009). https://doi.org/10.1007/BF02985802
14. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735
15. Jain, A., Sudheer, K., Srinivasulu, S.: Identification of physical processes inherent in artificial neural network rainfall runoff models. Hydrol. Process. **18**(3), 571–581 (2004)
16. Kahn, J.R., Vásquez, W.F., de Rezende, C.E.: Choice modeling of system-wide or large scale environmental change in a developing country context: lessons from the Paraíba do Sul river. Sci. Total Environ. **598**, 488–496 (2017)
17. Khair, A.F., Awang, M.K., Zakaraia, Z.A., Mazlan, M.: Daily streamflow prediction on time series forecasting. J. Theoret. Appl. Inf. Technol. **95**(4), 804 (2017)
18. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
19. Kohavi, R., et al.: A study of cross-validation and bootstrap for accuracy estimation and model selection. IJCAI **14**, 1137–1145 (1995)
20. Kratzert, F., Klotz, D., Brenner, C., Schulz, K., et al.: Rainfall-runoff modelling using long short-term memory (LSTM) networks (2018)
21. Krishna, B., Rao, Y.S., Nayak, P.: Time series modeling of river flow using wavelet neural networks. J. Water Resour. Prot. **3**(01), 50 (2011)
22. Laptev, N., Yosinski, J., Li, L.E., Smyl, S.: Time-series extreme event forecasting with neural networks at uber. In: International Conference on Machine Learning, pp. 1–5, no. 34 (2017)
23. Miguens, F.C., de Oliveira, M.L., de Oliveira Ferreira, A., Barbosa, L.R., de Melo, E.J.T., de Carvalho, C.E.V.: Structural and elemental analysis of bottom sediments from the Paraíba do Sul River (SE, Brazil) by analytical microscopy. J. South American Earth Sci. **66**, 82–96 (2016)
24. Mikolov, T., Karafiát, M., Burget, L., Černockỳ, J., Khudanpur, S.: Recurrent neural network based language model. In: Eleventh Annual Conference of the International Speech Communication Association (2010)
25. Ng, A., Katanforoosh, K., Mourri, Y.: Sequence models. Deeplearning. AI on Coursera (2018)
26. Patel, S.S., Ramachandran, P.: A comparison of machine learning techniques for modeling river flow time series: the case of upper cauvery river basin. Water Resour. Manag. **29**(2), 589–602 (2015)
27. Pena, E.H.M., de Assis, M.V.O., Proença, M.L.: Anomaly detection using forecasting methods ARIMA and HWDS. In: 2013 32nd International Conference of the Chilean Computer Science Society (SCCC), pp. 63–66 (2013). https://doi.org/10.1109/SCCC.2013.18
28. Salomão, M., Molisani, M., Ovalle, A., Rezende, C., Lacerda, L., Carvalho, C.: Particulate heavy metal transport in the lower Paraíba do Sul river basin, Southeastern, Brazil. Hydrol. Process. **15**(4), 587–593 (2001)
29. Shafaei, M., Kisi, O.: Predicting river daily flow using wavelet-artificial neural networks based on regression analyses in comparison with artificial neural networks and support vector machine models. Neural Comput. Appl. **28**(1), 15–28 (2017)
30. da Silva, I.N., Cagnon, J.Â., Saggioro, N.J.: Recurrent neural network based approach for solving groundwater hydrology problems. In: Artificial Neural Networks-Architectures and Applications. IntechOpen (2013)

31. Sobral, B.S., et al.: Drought characterization for the state of Rio de Janeiro based on the annual SPI index: trends, statistical tests and its relation with ENSO. Atmos. Res. **220**, 141–154 (2019)
32. Trento, A., Vinzón, S.: Experimental modelling of flocculation processes-the case of Paraiba do Sul Estuary. Int. J. Sedim. Res. **29**(3), 378–390 (2014)
33. Valipour, M., Banihabib, M.E., Behbahani, S.M.R.: Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir. J. Hydrol. **476**, 433–441 (2013)
34. Vásquez, W.F., de Rezende, C.E.: Willingness to pay for the restoration of the Paraíba do Sul River: a contingent valuation study from Brazil. Ecohydrol. Hydrobiol. (2018)
35. 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 (2015)
36. Zhang, J., Zhu, Y., Zhang, X., Ye, M., Yang, J.: Developing a long short-term memory (LSTM) based model for predicting water table depth in agricultural areas. J. Hydrol. **561**, 918–929 (2018)