





Multitask Learning for Predicting Natural Flows: A Case Study at Paraíba do Sul River

Gabriel Dias Abreu¹(✉)() , Leticia F. Pires¹(✉)() ,
Luciana C. D. Campos¹(✉)() , and Leonardo Goliatt²(✉)()

¹ Department of Computer Science, Federal University of Juiz de Fora,
Juiz de Fora, MG, Brazil

{gabrieldiasabreu, leticial, luciana.campos}@ice.ufjf.br

² Computational Modeling Program, Federal University of Juiz de Fora,
Juiz de Fora, Brazil

leonardo.goliatt@ufjf.edu.br

Abstract. Forecasting the flow of rivers is essential for maintaining social well-being since their waters provide water and energy resources and cause serious tragedies such as floods and droughts. In this way, predicting long-term flow at measuring stations in a watershed with reasonable accuracy contributes to solving a range of problems that affect society and resource management. The present work proposes the MultiTask-LSTM model that combines the recurring model of Deep Learning LSTM with the transfer of learning MultiTask Learning, to predict and share information acquired along the hydrographic basin of Paraíba do Sul river. This method is robust for missing and noisy data, which are common problems in inflow time series. In the present work, we applied all 45 measurement stations' series located along the Paraíba do Sul River basin in the MultiTask-LSTM model for forecasting the set of these 45 series, combining each time series's learning in a single model. To confirm the MultiTask-LSTM model's robustness, we compared its predictions' results with the results obtained by the LSTM models applied to each isolated series, given that the LSTM presents good time series forecast results in the literature. In order to deal with missing data, we used techniques to impute missing data across all series to predict the 45 series of measurement stations alone with LSTM models. The experiments use three different forms of missing data imputation: the series' median, the ARIMA method, and the average of the months' days. We used these same series with imputing data in the MultiTask-LSTM model to make the comparison. This paper achieved better forecast results showing that MultiTask-LSTM is a robust model to missing and noisy data.

Keywords: Multitask Learning · Forecasting hydrological time series · Deep learning · Long short-term memory (LSTM) · Paraíba do Sul River · Brazil

1 Introduction

The flow forecast is necessary due to the dependence and fixation of societies around river basins throughout history. It is fundamental for the civilization to maintain its essential activities, such as agriculture, livestock, basic sanitation, hydroelectric power generation, industry, and tourism. Keeping water available implies developing techniques to identify and predict the behavior of these basins. Besides, it is possible to avoid tragedies such as those resulting from floods, droughts, rupture of dams, and disease vectors [14]. From a current perspective of society, the improvement of these techniques is in line with the water resources' growth management and environmental preservation. It is negatively impacted by the accelerated urban expansion, enabling sustainable development and enabling decision-making and long-term risk planning competent bodies [10].

Historical records contained in time series of water phenomena are often costly and difficult to measure, in addition to presenting noises and missing data, which impairs the performance of forecasting these time series [6]. The case study of this work, the river basin's Paraíba do Sul, has 45 flow measurement stations with many missing data in all stations resulting from the station shutdown or the like activities. In addition to hydro-geomorphological modifications or even failures in sensors that result in noise in the time series.

The hydro-geomorphological variables present in a basin present correlated variations temporally and also spatially. That indicates possible events, such as changes in the records measured by an upstream flow measurement station, which influence the forecast of the downstream measurement station¹. Therefore, it is necessary to consider these phenomena to improve predictive capacity. For example, if a dam is installed in a river basin region, the entire flow downstream of that dam will be affected, so the time series forecasts of stations downstream from the dam need to consider this phenomenon.

The flow time series is susceptible to exogenous and uncertain factors, such as the measuring station's maintenance, probably because of measurement failures in sensors, which require its shutdown. Also, the relationships present in the time series distributed along the river basin, when not appropriately used, constitute a reneged potential for forecasting and wasting resources spent on flow measurements. Therefore, forecasting with robust methods for missing data and noise inflow time series is necessary.

MultiTask Learning is an approach to inductive learning transfer that increases generalization using information from related tasks. This is done by learning in parallel using a shared representation which can help to improve the learning of the others as defined in [4]. The MultiTask Learning method can be resilient to missing and noisy data since it considers the temporal and spatial relationships present in the river basin's flow time series. As a result,

¹ Downstream is the side where is directed the water flow and upstream is the part where the river is born. So, the mouth or outfall of a river is the most downstream point of this river, and the source is its most upstream point.

missing data or noise that would impair the model's performance has its negative effect diminished by the relationships present in the data, combining each time series's learning in a single model. The learning transfer method Multi-Task Learning still captures information implicit in the relationships between all flow time series along the river basin, providing better use of the available data concerning the forecast models' application separately in each measuring station. The motivation of this work consists of combining these characteristics of the transfer of learning MultiTask Learning with the LSTM model of recurrent neural networks.

The literature presents promising results in several applications. Jin and Sun [7] showed that multi-task learning (MTL) has the potential to improve generalization by transferring information in training signals of extra tasks. Ye and Dai [15] developed a multi-task learning algorithm, called the MultiTL-KELM, for multi-step-ahead time series prediction. MultiTL-KELM regards predictions of different horizons as different tasks. Knowledge from one task can benefit others, enabling it to explore the relatedness among horizons. Zhao and collaborators [17] introduced a multi-task learning framework that combines the tasks of self-supervised learning and scene classification. The proposed multi-task learning framework empowers a deep neural network to learn more discriminative features without increasing the parameters. The experimental results show that the proposed method can improve the accuracy of remote sensing scene classification. Cao et al. [3] proposed a deep learning model based on LSTM for time series prediction in wireless communication, employing multi-task learning to improve prediction accuracy. Through experiments on several real datasets, the authors showed that the proposed model is effective, and it outperforms other prediction methods.

The prediction of flow time series is widely used for the planning and management of water resources, as evidenced by the work in [13]. This paper presents the classic models such as ARIMA and Linear Regression, which are unable to capture the non-stationarity and non-linearity of the hydrological time series. This study also points to the growth of attention given to data-driven models such as neural networks that progress in predicting non-linear time series, capturing water time series's complexity. Aghelpour and Varshavian, [1] compare two stochastic and three artificial intelligence (AI) models in modeling and predicting the daily flow of a river. The results showed that the accuracy of AI models was higher than stochastic ones, and the Group Method of Data Handling (GMDH) and Multilayer Perceptron (MLP) produced the best validation performance among the AI models.

In comparison to several hydrological models, deep learning has made significant advances in methodologies and practical applications in recent years, which have greatly expanded the number and type of problems that neural networks can solve. One of the five most popular deep learning architectures is the long short-term memory (LSTM) network, which is widely applied for predicting time series [11]. LSTM is a specific recurrent neural network architecture that can learn long-term temporal dependencies and be robust to noise. This feature

makes it efficient in water resource forecasting problems as explored in the works at [9], which showed the LSTM model as an alternative to complex models. Such models can include prior knowledge about inflows' behavior and the study at [16] which showed LSTM's ability to predict water depth for long-term irrigation, thereby contributing to water management for irrigation. However, both works clarify the need for a considerable amount of data for LSTM to present satisfactory results.

The Paraíba do Sul River basin is of great importance for Brazilian economic development and supplies 32 million people [8]. This basin has 45 measurement stations whose captured time series have missing and noisy data, so forecasting this basin's flow is difficult. The work on [2] showed the efficiency of the LSTM model for the flow forecast in the Paraíba do Sul River basin compared to other classic models such as ARIMA and also pointed out the importance of the long flow forecast in this basin. This work used a subset of 4 of the 45 flow measurement stations on the Paraíba do Sul River.

To applying a Machine Learning technique to forecast time series, it is common to optimize an error measure by training a single forecast model of the desired time series. However, it is sometimes necessary to explore latent information from related series to improve forecasting performance, resulting in a learning paradigm known as Multi-Task Learning (MTL). According to Dorado-Moreno et al. [5] the high computational capacity of deep neural networks (DNN) can be combined with the improved generalization performance of MTL, designing independent output layers for each series and including a shared representation for them. The work of Shireen and collaborators [12] showed that models using MTL could capture information from several time-series simultaneously, with robustness to missing data and noise, making inferences about all historical data and their relationships within the scope photovoltaic panels.

This work proposes a robust forecasting model for missing and noisy data to make long-term flow predictions from information present in the time series of measuring stations located along a hydrographic basin. We have used the time series of measuring stations located along the Paraíba do Sul river basin as a case study for this work. The proposed model combines Deep Learning techniques, such as LSTM, with the transfer of learning MultiTask Learning - MTL, to take advantage of the implicit relationships between the time series of each measurement station, making the model robust to missing and noisy data to improve forecast performance.

2 Materials and Methods

2.1 Study Area and Data

The set of series used in this work consists of daily records collected, from 1935 to 2016, at 45 flow measurement stations along the Paraíba do Sul River basin, provided by the National Water Agency (ANA)². Some measurement stations

² www.ana.gov.br.

present missing or noisy data in their collected historical series, as can be seen in Fig. 1. The missing data in the series come from failures of sensors present in the measuring station or similar problems, which resulted in their shutdown for maintenance. In red are non missing data from a measurement station.

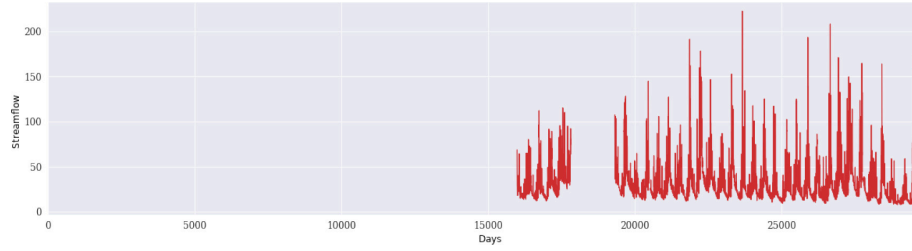


Fig. 1. Streamflow time series with missing data.

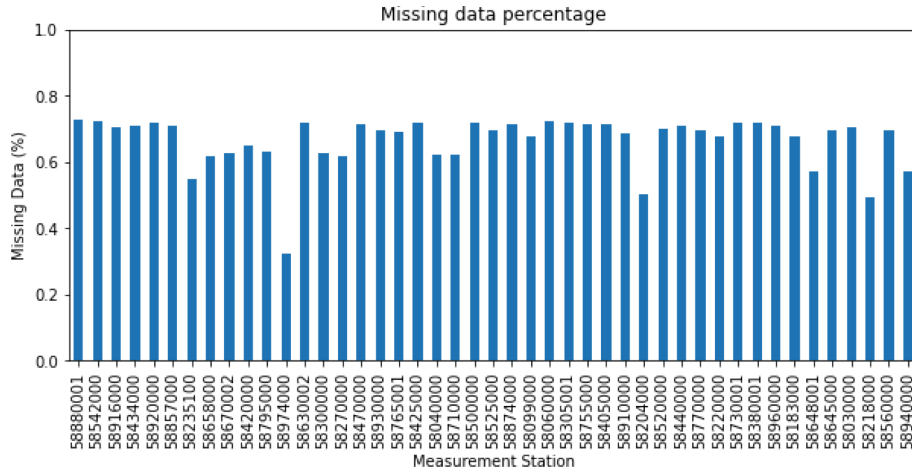


Fig. 2. Percentage of missing data per measurement station.

Missing data and noise are the problem for the time series' prediction since noise imply errors in learning the time series's behavior. On the other hand, missing data inhibits the model from understanding what happened when the data was unwilling. They, therefore, affect the continuity of the model forecast.

Figure 2 shows the number of records missing in the series of flow measurement stations along the basin, whether due to shutdown, maintenance, or defects present in the measurement stations in some period. The series' median, the ARIMA method, and the average of the months' days were some data imputations techniques to treat the missing data in these work's series. Simultaneously,

the MTL-LSTM model, which combines two robust techniques for dealing with noisy data, MultiTask Learning and LSTM, was used to deal with noises. As the imputing values' process in the missing data creates noises, the learning characteristics were from the correlation of the imputed time series in the MTL-LSTM model.

2.2 Streamflow Estimation Model

The experiments were carried out with the historical series' set of 45 flow measurement stations distributed along the Paraíba do Sul river basin to compare the forecast made by the MTL-LSTM models with the LSTM models trained with each isolated series.

As shown in Fig. 3, the E time series are provided as input to the model. They are divided into rolling windows of size j and steps of size 1. Each step of these time series is concatenated with the E measuring station, forming a E rows matrix and j columns and a y vector with size E . These data are then provided to the LSTM, which learns to predict the time series's future behavior.

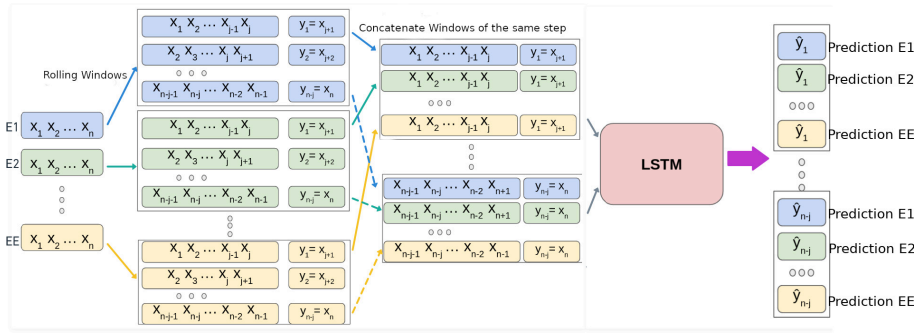


Fig. 3. MultiTask-LSTM model

The experiments were retrieved in the Google Colab³ environment with 12 GB of RAM in GPUs using the Keras⁴, NumPy⁵ and Tensorflow libraries⁶ in Python. All results were chosen about the average of 30 runs. The MAPE metric was chosen to compare the results, defined by the Eq. 1:

$$MAPE = \frac{100\%}{n} \sum_{t=1}^n \left| \frac{A_t - F_t}{A_t} \right| \quad (1)$$

³ colab.research.google.com.

⁴ keras.io.

⁵ numpy.org.

⁶ www.tensorflow.org.

where A_t is the historical time series value in time t , F_t is the value predicted in time t , and n is the size of the time series.

The LSTM applied in the MTL-LSTM model had hyper-parameters as suggested by Campos et al. [2]. These hyperparameters were used to build a single-task learning LSTM (STL-LSTM) to separately model each time series collected on the 45 measurement stations.

The MTL-LSTM model uses 14-day windows as in the work of Campos et al. [2], with 45 reference stations and is written as:

$$\begin{aligned} Q_{1,t+14} &= F(Q_{1,t}, Q_{1,t-1}, \dots, Q_{1,t-13}) \\ Q_{2,t+14} &= F(Q_{2,t}, Q_{2,t-1}, \dots, Q_{2,t-13}) \\ &\vdots \\ Q_{k,t+14} &= F(Q_{k,t}, Q_{k,t-1}, \dots, Q_{k,t-13}) \\ &\vdots \\ Q_{45,t+14} &= F(Q_{45,t}, Q_{45,t-1}, \dots, Q_{45,t-13}) \end{aligned}$$

where $Q_{k,t+14}$ is the streamflow at station k predicted 14d ahead.

We trained the model using a training set with the first 75% data of the time series, and the 10% of the followed data in the validation set to verify the hyperparameters, and the last 15% of data for the test set. Each experiment was performed 30 times, from which we calculated the average of the MAPE metric to assess the final performance of the model.

3 Computational Experiments

Figure 4 shows us that the MTL-LSTM model performs considerably better than the LSTM model when the median metric was applied in the imputation of missing data in the times series presented to the models, except for the station 58218000. The results evidence the MTL-LSTM model's capacity to learn hydro-geomorphological relations in the basin, ignoring the noise added by a constant median imputation.

When ARIMA or Mean of days per month are applied to impute missing values as in the Figs. 5 and 6, MTL-LSTM performs considerable better than LSTM in all measurement stations. This behavior shows the robustness of the MTL-LSTM model to learn series relations in the basin when data is more accurate with two imputation methods that preserve the seasonality and variability of the time series. This behavior indicates that MTL-LSTM would perform better than LSTM with no missing data.

The Table 1 summarizes the results found in the experiments. We can observe that the MultiTask-LSTM model obtains averaged percentage errors around half of the errors achieved with the individual LSTM models. Note that while the LSTM models achieved percentage errors above 40%, the MultiTask-LSTM model achieved MAPEs below 22%. As shown in Fig. 7, the MultiTask-LSTM has the advantage of having your training time faster as it places all flow measurement stations in the same model. On the other hand, the model containing only LSTM is considerably slower to train each time series separately.

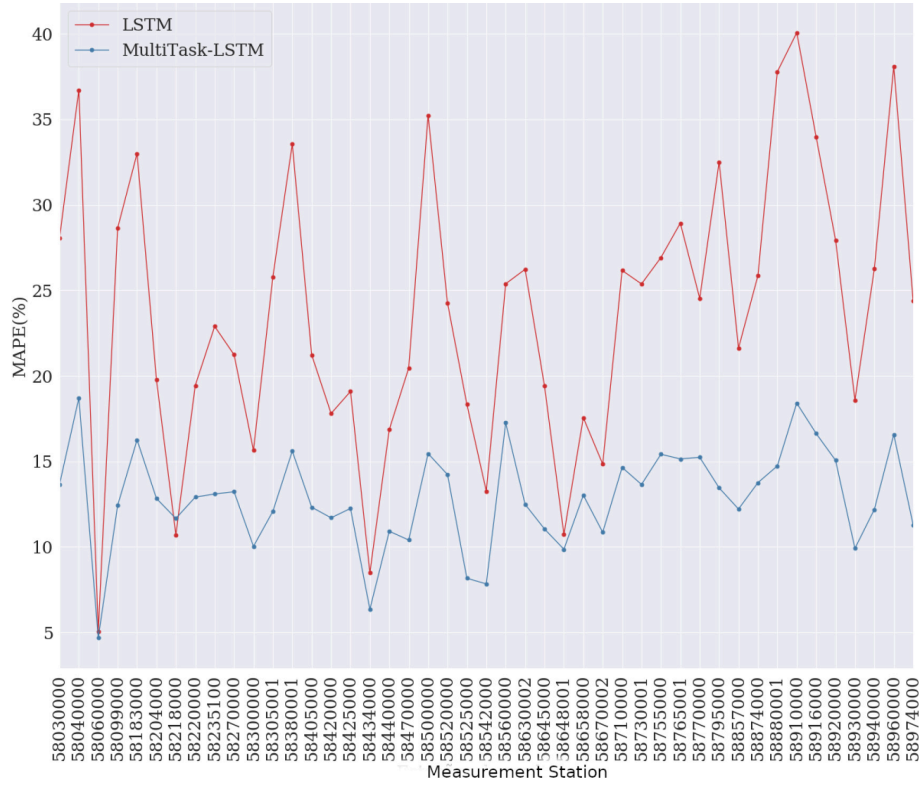


Fig. 4. MultiTask-LSTM and LSTM comparison with median missing data imputation

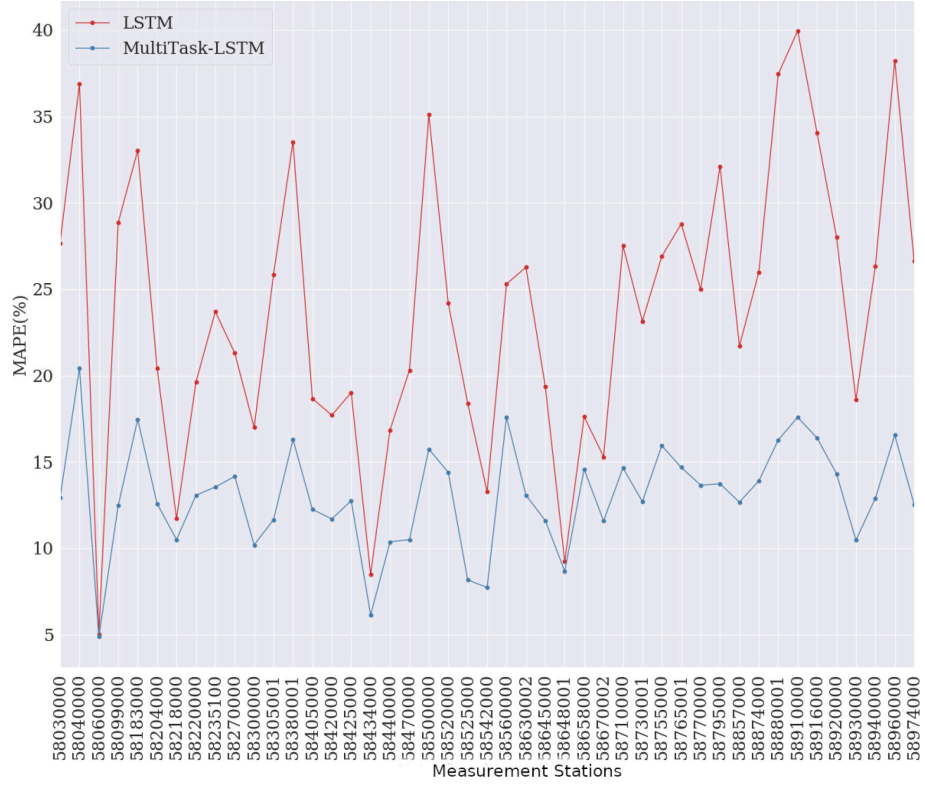


Fig. 5. MultiTask-LSTM and LSTM comparison with ARIMA missing data imputation

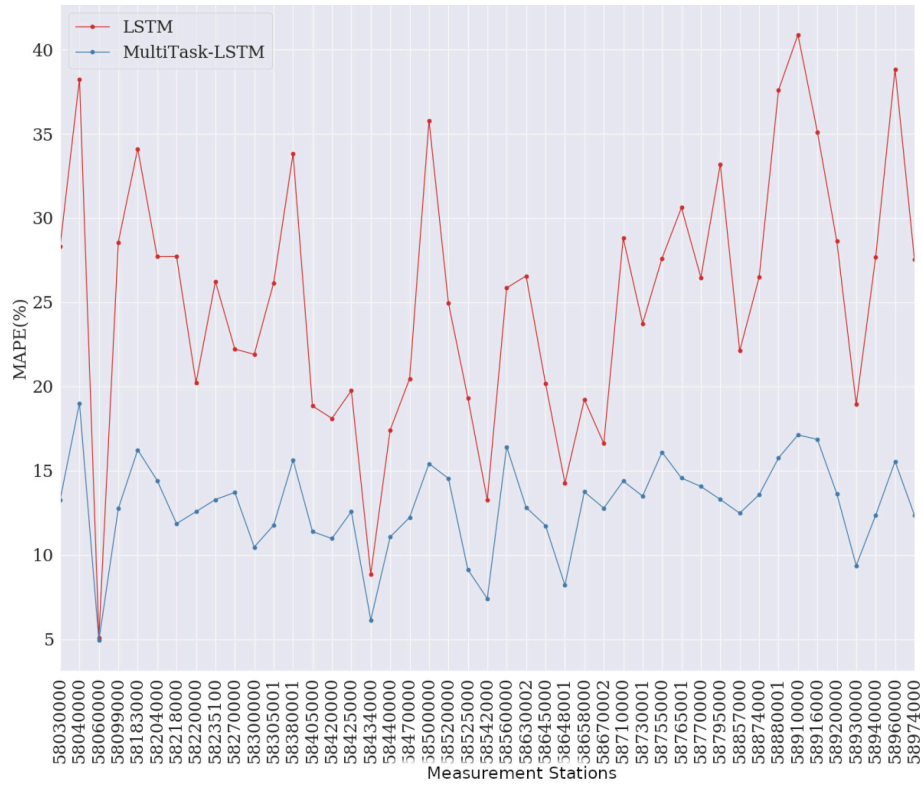


Fig. 6. MultiTask-LSTM and LSTM comparison with mean of days per month missing data imputation

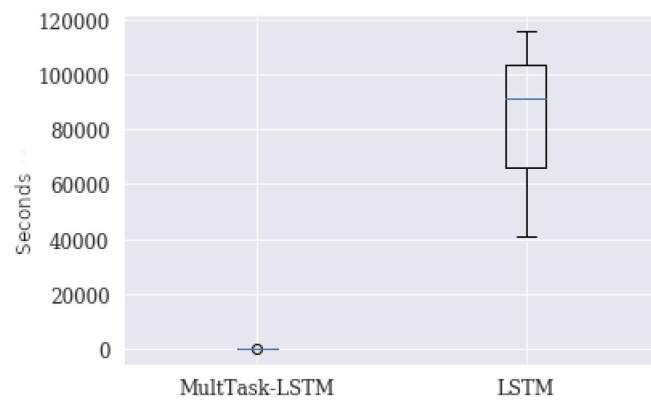


Fig. 7. Time comparison between MultiTask-LSTM and LSTM

Table 1. MAPE's mean for each streamflow measurement station by imputation method and model.

Stations (gauges)	ARIMA imputation		Mean imputation		Median imputation	
	LSTM	MultiTask-LSTM	LSTM	MultiTask-LSTM	LSTM	MultiTask-LSTM
58030000	27.67	12.91	28.31	13.25	28.04	13.64
58040000	36.90	20.44	38.25	19.01	36.72	18.71
58060000	5.03	4.87	5.08	4.92	5.04	4.67
58099000	28.87	12.48	28.56	12.75	28.65	12.43
58183000	33.03	17.46	34.11	16.23	33.01	16.26
58204000	20.42	12.58	27.70	14.41	19.77	12.83
58218000	11.71	10.49	27.71	11.85	10.70	11.66
58220000	19.61	13.06	20.22	12.56	19.44	12.91
58235100	23.70	13.54	26.23	13.28	22.89	13.09
58270000	21.30	14.16	22.22	13.71	21.23	13.22
58300000	16.99	10.19	21.90	10.46	15.65	10.03
58305001	25.85	11.65	26.15	11.76	25.80	12.08
58380001	33.53	16.29	33.81	15.63	33.56	15.60
58405000	18.66	12.25	18.84	11.38	21.23	12.32
58420000	17.70	11.69	18.10	10.97	17.80	11.70
58425000	19.01	12.75	19.75	12.56	19.09	12.24
58434000	8.48	6.14	8.85	6.14	8.48	6.35
58440000	16.84	10.37	17.41	11.06	16.86	10.91
58470000	20.31	10.50	20.45	12.22	20.45	10.40
58500000	35.10	15.74	35.79	15.41	35.24	15.45
58520000	24.18	14.39	24.95	14.53	24.26	14.23
58525000	18.39	8.17	19.32	9.13	18.33	8.16
58542000	13.27	7.72	13.26	7.40	13.25	7.83
58560000	25.30	17.58	25.85	16.42	25.38	17.25
58630002	26.27	13.07	26.56	12.82	26.23	12.49
58645000	19.35	11.60	20.17	11.74	19.40	11.04
58648001	9.23	8.67	14.27	8.20	10.73	9.86
58658000	17.61	14.57	19.22	13.74	17.56	13.02
58670002	15.29	11.58	16.61	12.77	14.84	10.85
58710000	27.51	14.65	28.81	14.38	26.16	14.64
58730001	23.15	12.69	23.74	13.49	25.37	13.65
58755000	26.90	15.94	27.61	16.09	26.92	15.41
58765001	28.78	14.68	30.62	14.56	28.93	15.14
58770000	24.99	13.65	26.46	14.06	24.53	15.24
58795000	32.11	13.73	33.19	13.29	32.50	13.45
58857000	21.70	12.67	22.11	12.47	21.62	12.21
58874000	25.99	13.90	26.48	13.57	25.86	13.75
58880001	37.45	16.27	37.59	15.75	37.78	14.73
58910000	39.96	17.59	40.88	17.12	40.07	18.41
58916000	34.03	16.38	35.11	16.85	33.98	16.63
58920000	28.03	14.31	28.63	13.63	27.95	15.06
58930000	18.59	10.46	18.93	9.36	18.58	9.91
58940000	26.32	12.89	27.66	12.34	26.27	12.19
58960000	38.23	16.54	38.84	15.55	38.10	16.57
58974000	26.63	12.53	27.54	12.35	24.38	11.28

4 Conclusion

Flow forecasting in the river basin is an essential issue for well-being and social development. To ensure adequate environmental, social and economic conditions, the study of models that provide the improvement of long-term flow forecasting is necessary, especially in time series with a lot of missing data, noise, and hydrogeomorphological changes such as flow time series.

Using the MultiTask Learning technique together with the Deep Learning model, LSTM, allows absorbing the information present in the data of all the time series of the measuring stations of a basin. In other words, it reuses the knowledge learned in a time series of a measuring station in the learning of the other series of that basin. The Paraíba do Sul River Basin, located in Brazil, was used as a case study for this work. However, the model can be applied to forecasting other basins where multiple flow measurement stations collect data, especially if these measuring stations have time series with noise or missing data.

The study used three missing data imputation techniques to verify robustness against noisy data of the MTL-LSTM model. As can be seen in Figs. 4, 5 and 6 the MTL-LSTM model achieved considerably better percentage errors in all missing data imputation scenarios. The LSTM models were applied in long-term forecasts in each series of flow measurement stations located along the Paraíba do Sul river basin. The MTL-LSTM model also presented a shorter training time when compared to the LSTM models, as seen in Fig. 7.

The learning transfer approach present in the MTL-LSTM model allowed the improvement of long-term forecasts. Results from all measuring stations in the hydrographic basin demonstrated the robustness of the data imputation procedure, maintaining a stable performance with the different imputations.

References

1. Aghelpour, P., Varshavian, V.: Evaluation of stochastic and artificial intelligence models in modeling and predicting of river daily flow time series. *Stoch. Environ. Res. Risk Assess.* **34**, 33–50 (2020). <https://doi.org/10.1007/s00477-019-01761-4>
2. Campos, L.C.D., Goliatt da Fonseca, L., Fonseca, T.L., de Abreu, G.D., Pires, L.F., Gorodetskaya, Y.: Short-term streamflow forecasting for Paraíba do Sul River using deep learning. In: Moura Oliveira, P., Novais, P., Reis, L.P. (eds.) *EPIA 2019. LNCS (LNAI)*, vol. 11804, pp. 507–518. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30241-2_43
3. Cao, K., Hu, T., Li, Z., Zhao, G., Qian, X.: Deep multi-task learning model for time series prediction in wireless communication. *Phys. Commun.* **44**, 101251 (2021)
4. Caruana, R.: Multitask learning. *Mach. Learn.* **28**(1), 41–75 (1997)
5. Dorado-Moreno, M., et al.: Multi-task learning for the prediction of wind power ramp events with deep neural networks. *Neural Netw.* **123**, 401–411 (2020)
6. Herschy, R.W.: *Streamflow Measurement*. CRC Press, Boca Raton (2014)
7. Jin, F., Sun, S.: Neural network multitask learning for traffic flow forecasting. *CoRR* abs/1712.08862 (2017). <http://arxiv.org/abs/1712.08862>
8. Kelman, J.: Water supply to the two largest Brazilian metropolitan regions. *Aquatic Procedia* **5**, 13–21 (2015). At the Confluence Selection from the 2014 World Water Week in Stockholm

9. Kratzert, F., Klotz, D., Brenner, C., Schulz, K., Herrnegger, M.: Rainfall–runoff modelling using long short-term memory (LSTM) networks. *Hydrol. Earth Syst. Sci.* **22**(11), 6005–6022 (2018). <https://doi.org/10.5194/hess-22-6005-2018>
10. Rezende, O.M., Miguez, M.G., Veról, A.P.: Manejo de águas urbanas e sua relação com o desenvolvimento urbano em bases sustentáveis integradas: estudo de caso dos Rios Pilar-Calombé, em Duque de Caxias/RJ. *Revista Brasileira de Recursos Hídricos* **18**(2), 149–163 (2013)
11. Sherstinsky, A.: Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D Nonlinear Phenomena* **404**, 132306 (2020)
12. Shireen, T., Shao, C., Wang, H., Li, J., Zhang, X., Li, M.: Iterative multi-task learning for time-series modeling of solar panel PV outputs. *Appl. Energy* **212**, 654–662 (2018)
13. 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)
14. Yassuda, E.R.: Gestão de recursos hídricos: fundamentos e aspectos institucionais. *Revista de Administração pública* **27**(2), 5–18 (1993)
15. Ye, R., Dai, Q.: Multitl-KELM: a multi-task learning algorithm for multi-step-ahead time series prediction. *Appl. Soft Comput.* **79**, 227–253 (2019)
16. 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)
17. Zhao, Z., Luo, Z., Li, J., Chen, C., Piao, Y.: When self-supervised learning meets scene classification: remote sensing scene classification based on a multitask learning framework. *Remote Sens.* **12**(20), 3276 (2020)