Does Deep Learning Advance Hourly Runoff Predictions?

Georgy Ayzel

[1] Institute for Environmental Sciences and Geography, University of Potsdam, Potsdam, Germany,
ayzel@uni-potsdam.de

Abstract

Timely and accurate flash flood predictions are crucial for early warning of this costly and deadly natural hazard. There are many modeling approaches tackling hourly runoff formation mechanisms and utilizing different levels of computational complexity and requirements to an input data. However, to keep an efficient trade-off between speed, data availability, robustness and interpretability, the focus of operational runoff forecasting systems development often shifts to conceptual hydrological models. However, the developing field of deep learning provides us an opportunity to advance flash flood predictions using new data-driven technologies. In the present study, we extensively search for optimal structure and parameters and examine the predictive performance of the Long Short-Term Memory (LSTM) model for continuous runoff predictions at an hourly time step and compare results with the process-based hydrological model. Results highlight that the LSTM model provides reliable hourly runoff predictions. Yet, it demonstrates only comparable efficiency with a conceptual hydrological model, but with a significant computational overhead.

1 Introduction

Runoff predictions have a key importance in the field of hydrology for both practical and theoretical perspectives. From the theoretical side, runoff predictions accumulate and concentrate the knowledge about runoff formation processes at different spatiotemporal scales [1]. From the practical side, they create a basis for water management, an assessment of available water resources regarding different scenarios of a future climate and a development of early warning systems of natural hazards [2, 3]. Propagation of emerging technological breakthroughs and investigation of their impact on runoff predictions effectiveness and efficiency play a critical role in advancing the field of scientific methodology and hydrologic engineering [4, 5].

There is a myriad of hydrological models doing the best for runoff predictions [6]. Each model may relate to one of the three main types regarding system simplification [4, 7]: physically based, conceptual, and data-driven. Based on simplicity, high computational efficiency and low input data requirements, conceptual and data-driven models are the most widespread both in theoretical studies and practical applications [4, 5, 7]. While the development of conceptual models was the most active research field in the last decade, the intensively-emerging field of machine learning provides new techniques to be extensively evaluated as data-driven hydrological models for runoff predictions [4, 5].

Applications of ANNs for runoff predictions and forecasting are the most widespread across the field of data-driven model development in hydrology. The most prominent advantages of ANN which support that domination are their abilities to handle incomplete, noisy and ambiguous data and to learn and generalize complex systems of input-output relations [7, 8]. Feed-forward ANNs are of the most use in runoff modeling studies. According to the latest review paper [9], they share around 74% of used ANN architectures, but with significant decreasing trend in favor of using other architectures, in comparison with the share of feed-forward ANNs reported by Dawson and Wilby [7] of 98% as of 2001. One of these new architectures is the recurrent neural network (RNN) approach which takes into account the temporal nature of the input time series. Despite theoretical advances, there is no strong evidence that RNNs outperform feed-forward ANNs for runoff predictions: Kumar et al. [10] showed that RNNs better predict monthly runoff, but Hsu et al. [11] and Taver et al. [12] showed the opposite for daily predictions.

The Long Short-Term Memory (LSTM) network introduced by Hochreiter and Schmidhuber [13] is a new type of RNN architecture that overcomes the limited memory capacity of RNNs and allows learning of long-term input-

In: Sergey I. Smagin, Alexander A. Zatsarinnyy (eds.): V International Conference Information Technologies and High-Performance Computing (ITHPC-2019), Khabarovsk, Russia, 16-19 Sep, 2019, published at http://ceur-ws.org

output dependencies. Recent studies show a high potential of LSTM models in hydrological modeling: it was shown that LSTM models can outperform both conventional feed-forward ANNs and conceptual hydrological models [14, 15].

The rapid growth of studies that involve modern deep learning architectures in the field of hydrology is ongoing right now, and many prospective studies are emerging [16]. The objective of this study is to investigate the capability of LSTM model reproduce rainfall-runoff process at an hourly temporal resolution and highlight its predictive potential, advances and shortcomings over the state-of-the-art conceptual hydrological model.

2 Data and Methods

2.1 Study Site and Hydrometeorological Data

We use the Rimbaud River basin at Collobrières as the study site. It is located in the Maures highlands of southeastern France, close to the Mediterranean Sea coast and drains an area of 1.4 km². For a detailed description of the basin characteristics, such as elevation, vegetation, geology, climatology, and a hydrologic regime, please see the Supplement material of Thirel et al. [17] paper.

Hourly precipitation data (P) was obtained from three different sources: two nearby rain gauges located outside the basin and the SAFRAN reanalysis [18]. When precipitation observations at a specific time were available on both rain gauges, we took an average value of both, if one of two observations was missing we took the measured one. In the situation no observations from any of the gauges were available, we took a precipitation estimation from the nearest SAFRAN reanalysis grid cell. Hourly potential evapotranspiration data (PE) were obtained from the SAFRAN reanalysis which utilizes the Penman-Monteith formulation to calculate PE. Discharge data at hourly time resolution were obtained from Irstea, France. Data is available from 1 January 1968 to 31 December 2004 and has no missing values for meteorological forcing (P and PE) and around 1% missing values for discharge.

2.2 Hydrological Model

In the present study we use GR4H – the process-based, lumped, conceptual hydrological model, which was developed for runoff predictions at an hourly temporal resolution [19]. GR4H has four free parameters which are calibrated by optimizing the Nash-Sutcliffe efficiency criteria (NS, see Sect. 2.5) using the differential evolution algorithm [20]. To initialize the GR4H model states we use a warm-up period that copies the period under consideration.

2.3 LSTM Model

In the present study, the chosen LSTM model is the ANN constructed from two types of calculation layers – LSTM and fully-connected (Dense) – to map meteorological forcing (input layer) to runoff (output layer) data (Fig. 1). For a detailed description of the whole workflow of implementing LSTM networks for runoff predictions, we recommend referring to Kratzert et al. [14].

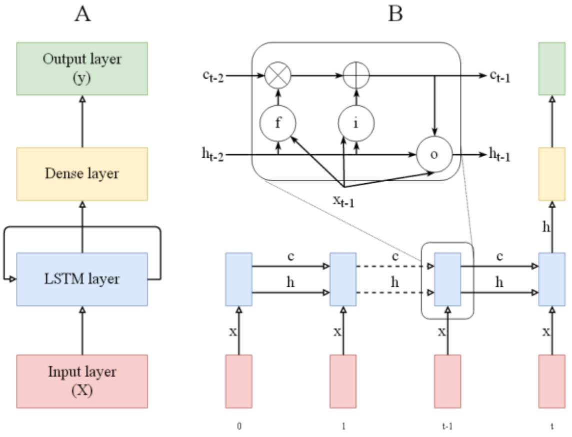

Figure 1: The LSTM model architecture

Objective guidelines for designing an appropriate structure of ANN and its hyperparameters for a specific modeling task remain elusive [7-9]. In most cases, researchers rely on trial-and-error techniques, and provide only the most effective and efficient architecture and hyperparameters, but do not provide information about their failures that could help researchers to avoid such failures, or investigate the reasons of the detected failures. In this study, we use grid search – a computationally-inefficient, but at the same time one of the most informative methods for finding a quasi-optimal setting of model architecture and related parameters in discrete space of parameters.

We constrained the hyperparameter space by considering only six important LSTM model hyperparameters, investigated separately for two variants of forcing data (precipitation and potential evapotranspiration; precipitation only) (see Fig. 2):

1. The input history length (hours), which is the length of the input sequence of meteorological forcing to be considered as a predictor for runoff. We use 25 candidates: from 48 to 1200 hours with a step of 48 hours.
2. The number of LSTM layers, which goes from one to three consecutive layers;
3. The number of LSTM units (# neurons), which is the length of hidden/cell states of LSTM layers. We use candidate values of 16, 32, and 64.
4. The regular dropout rate of the LSTM layer. We use two options: 0 and 0.2.
5. The recurrent dropout rate of the LSTM layer. We use two options: 0 and 0.5.
6. The number of Dense layers. We use the following candidates: one layer with one hidden neuron; two consecutive layers with 32 neutrons on the first and with one neuron on the second layer; three consecutive layers with 64, 32, and one neuron, correspondingly.

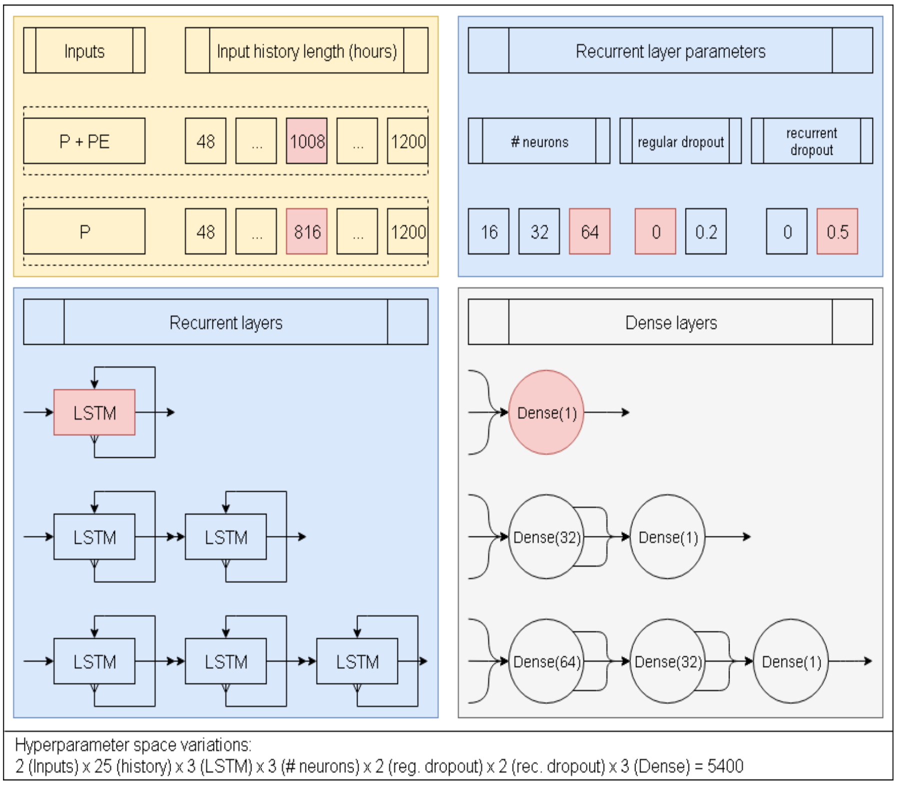

Figure 2: The LSTM model architecture. The optimal hyperparameters are shown in red boxes

For every candidate configuration, we trained the LSTM model and evaluate its corresponding performance on two independent periods: training and validation. For the validation period, we randomly reserve 4 years of available data (~12%), and the remaining data of 30 years has been used for the training period (88%). We used the early stopping technique to interrupt the training while there is no improvement for model performance on the validation period for 5 epochs (1 epoch is the time when each training sample has been entirely used for model optimization), and then save only the best model variant in terms of efficiency regarding the validation period. For model training, we use the RMSprop optimization algorithm with default hyperparameters and the mean squared error as a loss function.

2.4 Benchmark Setting

For the benchmark setting, we slightly adapted the standard split-sample test methodology proposed by Klemeš [21], which is widely used in many hydrological studies (Fig. 3). There are two major modifications:

1. We add an independent validation period (VP) which is isolated from the period for model calibration and testing.

2. In addition to a standard splitting of available data on two periods (P1 and P2) for model calibration (training) and testing, we also propose to augment these periods by the full period (FP).

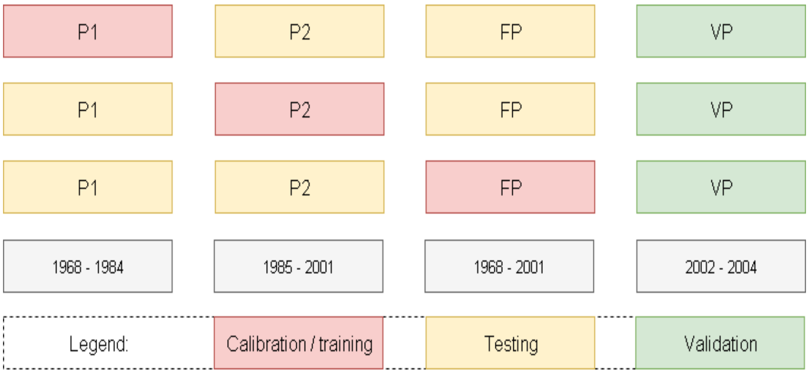

Figure 3: Modified standard split-sample test as a benchmark setting

The main goal of the proposed benchmark setting is to try to reach at some extent genericity for choosing a hydrological model of any type: for a fair comparison, we have to estimate GR4H and LSTM model performances under the same testing conditions.

2.5 Evaluation Metrics

The simulated runoff of the models was evaluated with widely-used criteria: the Nash-Sutcliffe efficiency (NS, [22]), the Kling-Gupta efficiency (KGE, [23]) and the absolute percentage of Bias [24]. NS and KGE are optimal with a value of 1 and are positively oriented, while Bias is optimal with a value of 0 and is negatively oriented.

3 Results and Discussion

3.1 Do We Need Deep Data-Driven Models for Runoff Predictions at Hourly Time Scale?

Interestingly, deep data-driven models might not be necessarily required. Results of the implemented extensive grid search procedure (Fig. 2) show that the development of a reliable LSTM model for hourly runoff predictions does not require deep architectures that utilize many sequential LSTM and Dense layers. In the context of runoff predictions, it is in our case a better strategy to build a shallow model with one, but wide, LSTM layer (with a high number of neurons) with consequent one and narrow (with the one neuron) Dense layer, and the high rate of recurrent regularization. Obtained results are consistent with the earlier study by Zhang et al. [15] who showed that a model with one LSTM layer outperforms a model with two sequential LSTM layers. Our results also underline the importance of using any structured algorithm for an extensive search for the best possible LSTM-based architecture and its corresponding hyperparameters in the defined parameter space, rather than poorly documented trial-and-error experiments [14, 15]. This search can not only improve the identification of suitable model structures, but also provide insights into modeling system features (e.g. relationships between available input data and model hyperparameters), and serves as a guidance for further studies in the field of machine learning in hydrology.

3.2 How Do GR4H and LSTM Models Perform?

Results reveal clear differences in efficiency patterns between GR4H and LSTMs (Fig. 4). For calibration and testing periods (P1, P2, FP) GR4H has a more uniform spread of efficiency metrics disregarding which data was used for model calibration and validation: NS ranges between 0.69 and 0.77, KGE ranges between 0.71 and 0.84 and Bias ranges between 0 and 14 percent. For LSTM_PPE, these intervals are 0.62-0.84 (NS), 0.65-0.9 (KGE) and 1.5-12 (Bias). For LSTM_P, these intervals are 0.55-0.8, 0.61-0.87 and 2.6-22, correspondingly. In contrast to GR4H, identified patterns in LSTMs efficiencies highlight the strong natural dependency of data-driven models efficiency and robustness to the training data (as also observed in [25]). Results also show the clear model outperformance on P1 (1968-1984) over P2 (1985-2001) period, that reflects the change of basin natural conditions driven by the severe wildfire in 1990, when over 84% of the basin was burnt [17].

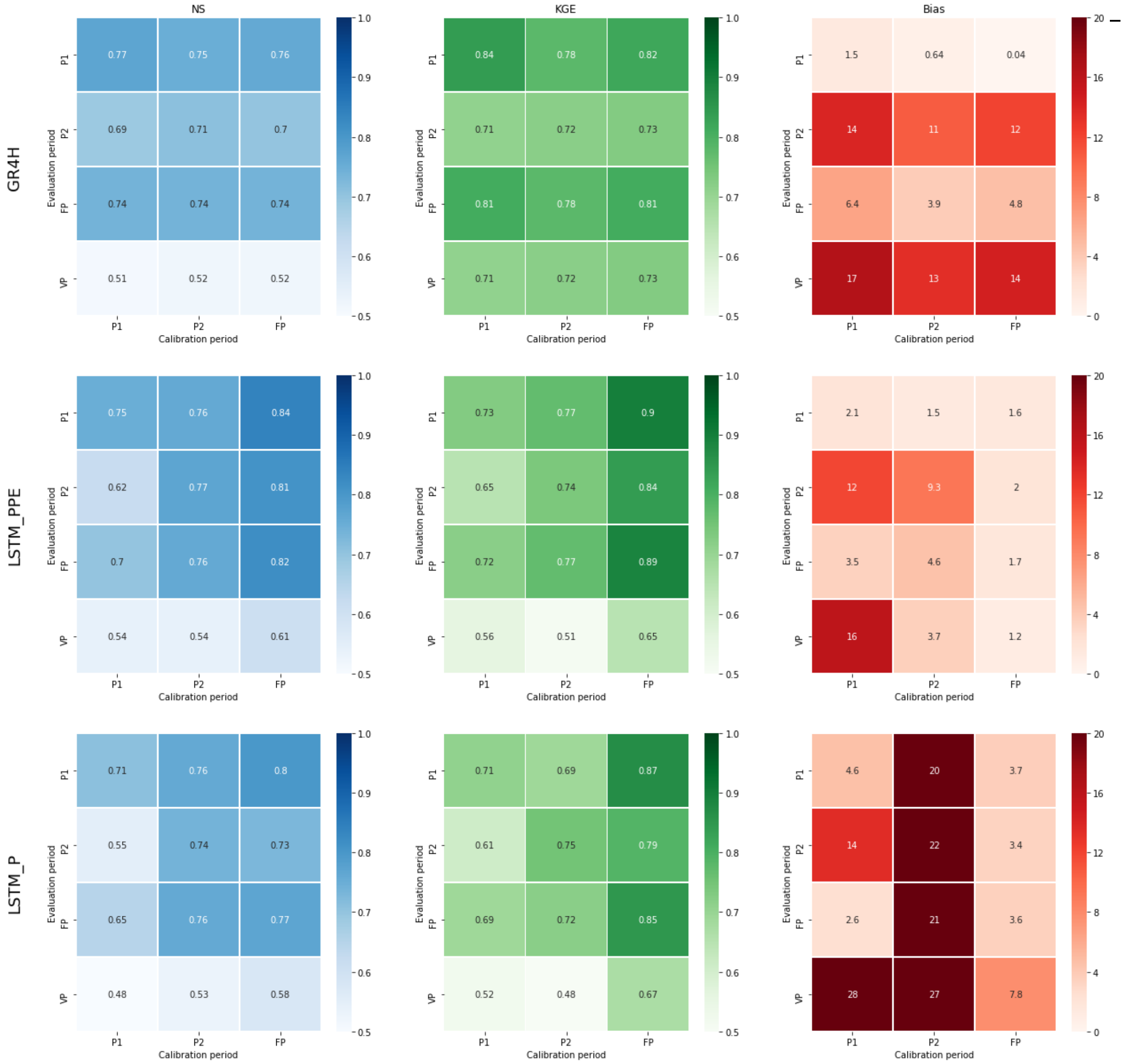

Figure 4: Model performance evaluation results

On the independent validation period (VP, 2002-2004) all models show a substantial drop of efficiency disregarding which data was used for model calibration/training. Both LSTM models trained on a full calibration period (FP, 1968-2001) outperform the GR4H model: LSTMs have a higher NS (0.61 and 0.58 for LSTMs in comparison with 0.52 for GR4H) and a lower Bias (1.2 and 7.8 for LSTMs in comparison with 14 for GR4H). However, an additional use of the KGE shows a reverse situation where GR4H outperforms both LSTMs (KGE is 0.73 for GR4H in comparison with 0.65 and 0.67 for LSTMs). Further decomposition of KGE into separate components that describe the contribution of correlation, variance, and bias clearly shows that the reason for this behavior is the higher contribution of variance component for both LSTMs. Visual inspection of simulated hydrographs shows that LSTMs do not capture small runoff peaks.

Results show that a split-sample test is a helpful tool for model efficiency and robustness analysis, but in case of implementing rainfall-runoff models for operational use in real setting it is better to use all available data for model calibration/training disregarding results of split-sample test: all models show better results when calibrated/trained on a full period. Thus, our findings also confirm the same conclusion which was provided in [26]. The gain of using a full period for model calibration/training is higher for data-driven (LSTM) models.

Using confined forcing (only precipitation time series) to drive LSTM models shows a good potential in the light of further integration of LSTM-based model into operational setting driven by the only precipitation data obtained from weather radars. Trained on a full period, the LSTM_P model shows comparable results with the LSTM_PPE model: the former has a lower NS and higher Bias, but also a lower variance. The lower variance of LSTM_P model shows its best performance for small runoff peaks (where runoff is under 1 mm/hour). This behavior proves higher and faster response of LSTM_P model to precipitation signal in comparison with the more inert LSTM_PPE model, which does additionally utilize potential evapotranspiration signal.

3.3 To What Extent Are Flash Floods Predictions Benefiting From Using LSTM Models?

Results do not provide a clear answer to that question. We highlight six events where the peak runoff is above 2 mm/hour for testing the models' efficiency for flash floods predictions on the validation period (Fig. 5). For three of them (Fig. 5, events B, C, D) LSTM_PPE shows better results than GR4H, for two events there is a tie (Fig. 5, events A, E), and only for one event (Fig. 5, event F) GR4H is better. Results of LSTM_P model is worse: it is better for runoff predictions than GR4H for one event only (Fig. 5, event B), there is almost no difference between models for two events (Fig. 5, events D, F), and GR4H outperforms LSTM_P for the remaining three events (Fig. 5, events A, C, E).

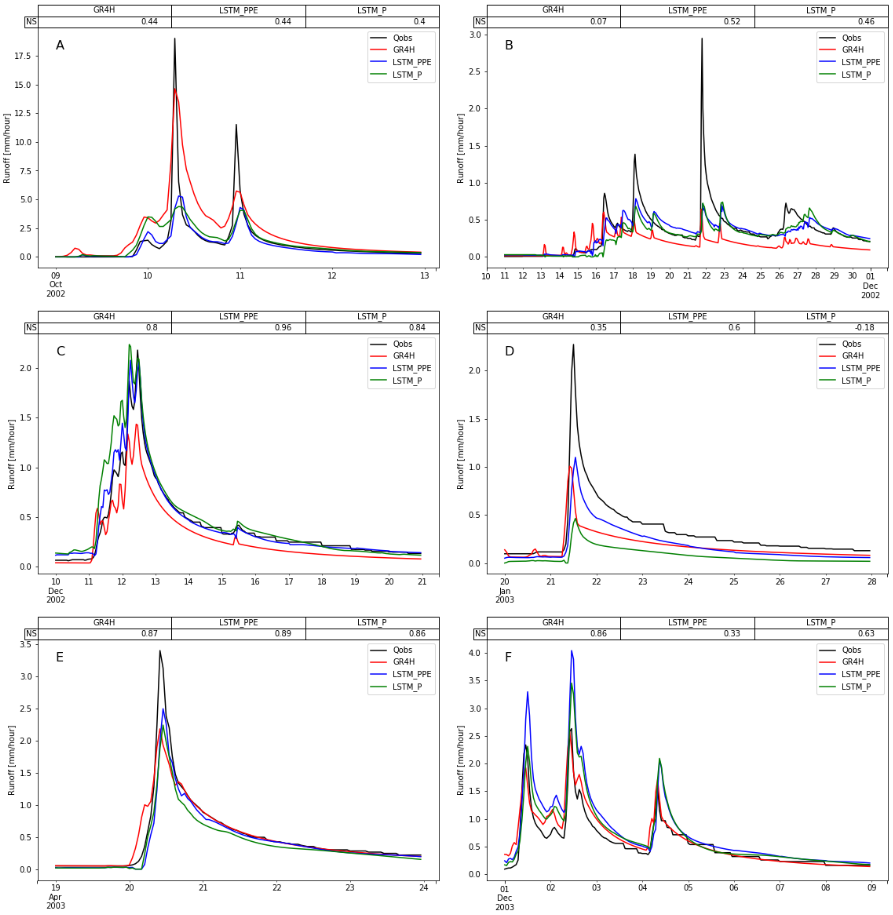

Figure 5: Simulated runoff timeseries for flash floods during the validation period

Flash floods predictions are not in the particular focus of this study, which strives to show the performance of different types of models in general. However, the LSTM_PPE model shows a great potential for flash floods predictions with results that are at least comparable with the process-based GR4H model. Unfortunately, for bigger flash floods (where the peak runoff is above 2 mm/hour), LSTM_P shows limited reliability. For the events with duration longer than one hour it is critical to include potential evapotranspiration time series in LSTM model forcing, which indirectly represents inert basin state. This finding contrasts with Toth and Brath results [25], which identified no improvement of accounting for potential evapotranspiration time series in comparison with the use of precipitation data alone. However, in contrast to Toth and Brath [25], we do not assimilate antecedent runoff observations, which can hold the most of valid information for runoff modeling due to the high runoff autocorrelation. In comparison with the hard-coded structure of GR4H, LSTM models have a potential to focus on a particular type of events using an oversampling technique – when during the training we feed to neural network not only one realization of the event per epoch but several identical realizations. This way, data-driven models acquire additional flexibility for runoff modeling studies.

4 Summary and Conclusions

One hydrological and two LSTM models were extensively evaluated for runoff predictions at an hourly temporal resolution on the small basin in France with the Mediterranean climate using an updated split-sample test and multiple efficiency criteria. The main findings of the present work can be summarized as the following:

1. Deep learning belongs to the groundbreaking technologies that are revolutionizing many fields of scientific research by setting up the new level of predictability [16]. However, at the moment its applications for runoff predictions show comparable results with conventional process-based hydrological models [14]. Our study suggest that new data-driven models (e.g. LSTM models) are not necessarily a new breakthrough in runoff modeling, but reliable members in the family of hydrological models with their own advantages and disadvantages.
2. In contrast to deep learning as the name of a group for the new data-driven methods, these models do not necessarily have to be deep. In our study, we show that shallower and more regularized models provide better results.
3. We show the critical importance of implementing two techniques for exploratory analysis of hydrological models that are often ignored in hydrological studies: a grid search and model performance assessment using multiple evaluation metrics. Grid search not only helps us to objectively identify the quasi-optimal LSTM models, but also reveals differences among optimal LSTM models, which have a hydrological meaning. Using multiple statistical criteria for assessing model performance helps not to miss important details in model behavior and identify their possible sources.
4. In contrast to GR4H, the performance of LSTM models strongly depends on training data. There are two following practical implications. In case little data of observed runoff is available for model calibration, it is better to use a process-based hydrological model (e.g. GR4H).
5. More calibration data – better performance on an independent period. That result was obtained disregarding the model we use for predictions and also has an important practical implication: if the final goal is providing modeling solutions to be implemented in operational use or decision making, it is a better strategy to use all the data available for optimizing proposed models.
6. We do not consistently advance flash floods predictions using LSTMs. Tackling this problem, we have to programmatically shift the model’s focus to flash floods using techniques for learning patterns from imbalanced datasets (e.g. oversampling).

A review of previous research studies in the field of data-driven hydrology leaves us feeling that we as a community are walking in a circle. Despite extensive guidelines how to organize research in a sustainable way and push it further [7-9], we still choose directions of least resistance and risk. The possible way to leave this flat circle is to synthesize dialog between two different research communities (from the side of hydrology, and from machine learning) and benefit from their common expertise, e.g. by running new model intercomparison projects that will pursue the most challenging research directions.

Acknowledgements

The present study was mainly funded by the Russian Foundation for Basic Research (RFBR), project no. 19-05-00087 A. The hydrological model development and evaluation part (Sect. 2.2, 3.2) was supported by the Russian Science Foundation, project no. 16-17-10039. Georgy Ayzel was financially supported by Geo.X, the Research Network for Geosciences in Berlin and Potsdam.

Georgy Ayzel would like to thank Météo-France for making available the meteorological SAFRAN data and Dr. Nathalie Folton (Irstea Aix-en-Provence) and Dr. Guillaume Thirel (Irstea) for providing the meteorological and discharge gauge data. Georgy Ayzel thanks Dr. Maik Heistermann and Dr. Guillaume Thirel for their helpful recommendations and positive criticism of this study.

Computational experiments have been carried out using GPU computation resources of the Machine Learning Group of the University of Potsdam (Potsdam, Germany) and the Shared Facility Center “Data Center of FEB RAS” (Khabarovsk, Russia).

References

1. Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., ... & Fenicia, F.: A decade of Predictions in Ungauged Basins (PUB)—a review. Hydrological sciences journal, 58(6), 1198-1255. (2013)

2. Montanari, A., Young, G., Savenije, H. H. G., Hughes, D., Wagener, T., Ren, L. L., ... & Blöschl, G.: “Panta Rhei—everything flows”: change in hydrology and society—the IAHS scientific decade 2013–2022. Hydrological Sciences Journal, 58(6), 1256-1275. (2013).

3. McMillan, H., Montanari, A., Cudennec, C., Savenije, H., Kreibich, H., Krueger, T., ... & Di Baldassarre, G.: Panta Rhei 2013–2015: global perspectives on hydrology, society and change. Hydrological sciences journal, 61(7), 1174-1191. (2016).

4. Solomatine, D. P., Ostfeld, A.: Data-driven modelling: some past experiences and new approaches. Journal of hydroinformatics, 10(1), 3-22. (2008).

5. Mount, N. J., Maier, H. R., Toth, E., Elshorbagy, A., Solomatine, D., Chang, F. J., Abrahart, R. J.: Data-driven modelling approaches for socio-hydrology: opportunities and challenges within the Panta Rhei Science Plan. Hydrological Sciences Journal, 61(7), 1192-1208. (2016).

6. Paniconi, C., Putti, M.: Physically based modeling in catchment hydrology at 50: Survey and outlook. Water Resources Research, 51(9), 7090-7129. (2015).

7. Dawson, C. W., Wilby, R. L.: Hydrological modelling using artificial neural networks. Progress in physical Geography, 25(1), 80-108. (2001).

8. Maier, H. R., Dandy, G. C.: Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environmental modelling & software, 15(1), 101-124. (2000).

9. Maier, H. R., Jain, A., Dandy, G. C., Sudheer, K. P.: Methods used for the development of neural networks for the prediction of water resource variables in river systems: current status and future directions. Environmental modelling & software, 25(8), 891-909. (2010).

10. Kumar, D. N., Raju, K. S., Sathish, T.: River flow forecasting using recurrent neural networks. Water resources management, 18(2), 143-161. (2004).

11. Hsu, K. L., Gupta, H. V., Gao, X., Sorooshian, S., Imam, B.: Self-organizing linear output map (SOLO): An artificial neural network suitable for hydrologic modeling and analysis. Water Resources Research, 38(12), 38-1. (2002).

12. Taver, V., Johannet, A., Borrell-Estupina, V., Pistre, S.: Feed-forward vs recurrent neural network models for non-stationarity modelling using data assimilation and adaptivity. Hydrological sciences journal, 60(7-8), 1242-1265. (2015).

13. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation, 9(8), 1735-1780. (1997).

14. Kratzert, F., Klotz, D., Brenner, C., Schulz, K., Herrnegger, M.: Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. Hydrology and Earth System Sciences, 22(11), 6005-6022. (2018).

15. 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. Journal of hydrology, 561, 918-929. (2018).

16. Shen, C., Laloy, E., Elshorbagy, A., Albert, A., Bales, J., Chang, F. J., ... & Fang, K. HESS Opinions: Incubating deep-learning-powered hydrologic science advances as a community. Hydrology and Earth System Sciences, 22(11). (2018).

17. Thirel, G., Andréassian, V., Perrin, C., Audouy, J. N., Berthet, L., Edwards, P., ... & Lindström, G. Hydrology under change: an evaluation protocol to investigate how hydrological models deal with changing catchments. Hydrological Sciences Journal, 60(7-8), 1184-1199. (2015).

18. Vidal, J. P., Martin, E., Franchistéguy, L., Baillon, M., Soubeyroux, J. M.: A 50-year high-resolution atmospheric reanalysis over France with the Safran system. International Journal of Climatology, 30(11), 1627-1644. (2010).

19. Mathevet, T.: Which rainfall-runoff model at the hourly time-step? Empirical development and intercomparison of rainfall runoff model on a large sample of watersheds (Doctoral dissertation, PhD thesis, ENGREF University, Paris, France). (2005).

20. Storn, R., Price, K. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. Journal of global optimization, 11(4), 341-359. (1997).

21. Klemeš, V. Operational testing of hydrological simulation models. Hydrological Sciences Journal, 31(1), 13-24. (1986).

22. Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I—A discussion of principles. Journal of hydrology, 10(3), 282-290.

23. Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. Journal of hydrology, 377(1-2), 80-91.

24. Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., & Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Transactions of the ASABE, 50(3), 885-900.

25. Toth, E., & Brath, A. (2007). Multistep ahead streamflow forecasting: Role of calibration data in conceptual and neural network modeling. Water Resources Research, 43(11).

26. Arsenault, R., Brissette, F., & Martel, J. L. (2018). The hazards of split-sample validation in hydrological model calibration. Journal of hydrology, 566, 346-362.

27. Paniconi, C., & Putti, M. (2015). Physically based modeling in catchment hydrology at 50: Survey and outlook. Water Resources Research, 51(9), 7090-7129.

28. Dawson, C. W., & Wilby, R. L. (2001). Hydrological modelling using artificial neural networks. Progress in physical Geography, 25(1), 80-108.