Operational discharge forecasts assimilating pre-trained Deep Learning models

Bob E. Saint Fleur, Eric Gaume, Florian Surmont, Nicolas Akil

Operational discharge forecasts assimilating pre-trained Deep Learning models

Bob E SAINT FLEUR[1], Eric GAUME[1], Florian SURMONT[1], Nicolas AKIL[2]

1- Universite Gustave Eiffel - GERS-EE
Allee des Ponts et Chaussees, 44344 Bouguenais - France

2- aQuasys Entreprise
2 rue de Nantes, 44710 Port-Saint-Pere- France

Abstract. Recent works have shown the predominance of Deep Learning models, particularly LSTM, over traditional rainfall-runoff models [1]. However, while operational hydrology requires accurate discharge forecasts, these models are predominantly designed for simulation. As a drawback, they are also limited of performing data assimilation or persistence analysis, which remain crucial for effective forecast and analysis. Therefore, we propose, using three DA techniques, to separately integrate the benchmark models LSTM [1] and SACSMA [2] into an orchestrator to provide forecasts. Using the CAMELS dataset [3], the DA technique's added-values are assessed comparing to original benchmarks on 3 levels of lead time. The results indicate significant improvement, and remain relevant to the classical conclusions on DA expectations.

1 Introduction

In operational hydrology, discharge forecast models remain essential. However, providing forecast remains a challenging task. The use of data assimilation (DA) techniques remains one of the best strategies to overcome these challenges, it is even more crucial for short-term discharge forecasts [4, 5]. These techniques, including persistency analysis [6], may be inappropriate on simulation models.

Through years, numerous hydrologic models have been developed. Based on assumption on the hydrological processes, the existing may be of physics-based, conceptual or empirical including data-driven type [7]. While data-driven models rely mainly on statistical relationships between rainfall and discharge, recent studies have stated their predominance over the classical approaches [1]. Dominated by the Deep Learning (DL) models, the most popular include Multilayer Perceptron (MLP), Long-Short-Term Memory (LSTM) [1] and Transformers [8]. They owe their success to the availability of large dataset, advances in computing capacity, powerful optimization algorithms (ADAM, SGD) and effective feature engineering. However, these models have been mainly designed for discharge simulation, and barely provide persistence analysis. Therefore we propose here to address these limitations using three DA techniques on pre-trained benchmark models, while assessing their improvement gains.

We use an MLP as a forecasting orchestrator of the benchmark works of [1] and [2] through three distinct DA techniques: (1) assimilate the recent discharge data; (2) assimilate the benchmark models simulation; (3) post-process

the benchmark model simulation errors. We perform direct discharge forecasting at 1-, 3-, and 7-day lead time, using weather forecast as either perfect data or sampled from the historical records.

2 Methodology and materials

2.1 Data presentation

This study uses the CAMELS dataset [3], particularly the Maurer's forcings of 531 basins on the 1989-2008 period. Discharge are from the USGS streamflow. The training/calibration and evaluation period cover respectively the 1989-2006 and 2006-2008 period. The ensemble forecasting data is sampled on a date-to-date basis on the historical record, and only on 56 basins.

2.2 Data assimilation techniques

Using discharge from both record and simulations (benchmarks), we perform DA through three approaches. These approaches are formulated as following:

- Prediction assimilating the recent discharge measures ($Q^o_{t:t-p}$) as in Eq.1.

$$\hat{Q}_{t+hp} = f(Q^o_{t:t-p}, X^f_{t+hp:t-n}, X_{t:t-n}) \tag{1}$$

- Prediction informed with other models simulation ($Q^s_{t+hp:t-p}$), as in Eq.2.

$$\hat{Q}_{t+hp} = f(Q^s_{t+hp:t-p}, Q^o_{t:t-p}, X^f_{t+hp:t-n}, X_{t:t-n}) \tag{2}$$

- Simulation error post-processing (**ePP**) within three steps: estimating the error (ε_t), then get predicted at the specified lead time $\hat{\varepsilon}_{t+hp}$, finally used to correct the simulation at the same lead time $\hat{Q}_{t+hp}$. See Eq.3.

$$\begin{aligned} \varepsilon_t &= Q^o_t - Q^s_t \\ \hat{\varepsilon}_{t+hp} &= f(\varepsilon_{t:t-p}, Q^o_{t:t-p}, X^f_{t+hp:t-n}, X_{t:t-n}) \\ \hat{Q}_{t+hp} &= Q^s_{t+hp} + \hat{\varepsilon}_{t+hp} \end{aligned} \tag{3}$$

Where, upper-script notations ^{f},^{o},^{s} indicate respectively meteorological forecasts, observed runoff, and assimilated data. Under-scripts t, hp, n and p stand for actual time step, forecast lead time, sequence length on forcing and assimilated data. Forcings (and discharge) are marked with an X (and Q) respectively.

2.3 Models and setup

We use the regional LSTM proposed in [1] and the SACSMA from [2] to only simulate the flow on the whole 1989 - 2008 period. For the forecasting orchestrator (MLP) configuration, only the hidden layer (HL), learning_rate (LR) and the activation function (F) have been optimized utilizing a grid search cross-validation. The dominant hyper-parameters are $HL : [120, 90], [120, 90, 60]$, $LR : [0.01, 0.001]$, $F : [ReLU, Tanh]$.

2.4 Evaluation criteria

Two categories of criteria are used: (1) deterministic, using he persistence criterion [6]; and (2) probabilistic distribution on the ensemble forecast, using graphical chart such as Talagrand diagram [9], Receiving Operational Characteristic (ROC) curve [10, 11], including its AUC score are used. The persistency is described in Eq. 4, and may be ranged in $(-\infty, 1]$ where 1 indicates the perfect model. If ≤ 0, the model is worse or as bad as a naive forecast.

$$PERS = 1 - \frac{\sum_{t=hp}^{T} \left(Q_t - \hat{Q}_t\right)^2}{\sum_{t=hp}^{T} \left(Q_t - Q_{t-hp}\right)^2} \tag{4}$$

3 Results

The findings are presented using exclusively graphics. Separate colors ans linestyles are used to distinguish between approaches and benchmarks. Analysis may be based on comparison between the original benchmarks and the rest, or between the baseline and the rest others, issuing the benefit of the DA techniques. The line styles are as following: baseline (MLP Simple) in black, SACSMA cases in blue to violet, LSTM cases in red to orange.

3.1 Deterministic analysis

The persistence values are plotted in Fig. 1. The number of basins is distributed on Y-axis, while the criterion is on X-axis. Unsurprisingly, it remains

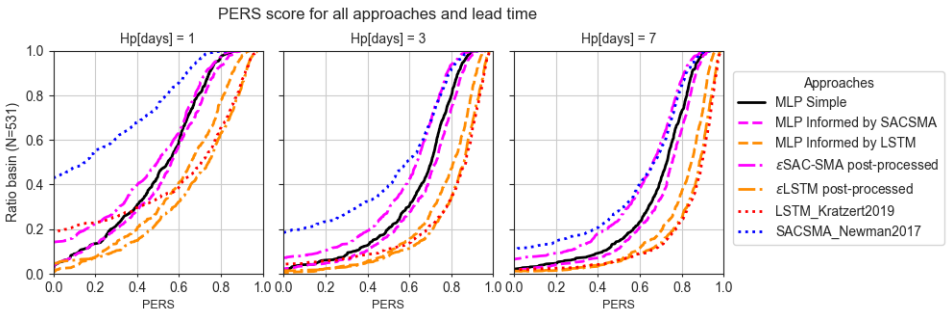

Fig. 1: Persistence (PERS) values

harder to forecast on shortest lead time. Almost 20% (or 40%) of the 531 tested basins issue a null persistence on the benchmarks (dotted). Using DA techniques, the baseline has issued a better performance, lowering this rate to 4%. Assimilating the benchmark models (dashed), the models issue is a systematic positive gain. When post-processing the simulation errors (dash-dotted), the performances shift in opposite directions between benchmarks and around the baseline; better for LSTM. However, when comparing the **ePP** cases to the

benchmarks themselves, the improvement is clearer, even though it remains tiny for the LSTM on the longer lead time.

3.2 Ensemble analysis

The ensemble analysis provides insight of the capacity of the models to provide reliable and accurate forecasts. It is expected to the DA techniques to help the models to deal with the high uncertainties linked to the weather forecasts data. When assimilating the target values, it has been proved in [4] that this it could weight around 50% of the information used by the models. While this may be a limitation of DA, it remains important given the poor quality of forecast, provided that the model is not naive.

3.2.1 Rank diagram

The reliability assessment can be done by the rank diagram, analyzing the uniformity with which the observation are forecasted over the time. It is then expected a flat-distribution for the reliable forecasts. Poor quality forecasts will issue either a Dom-shape (over dispersion), U-shape (under dispersion), Right- or Left-skewed (systematic under-estimation or over-estimation).

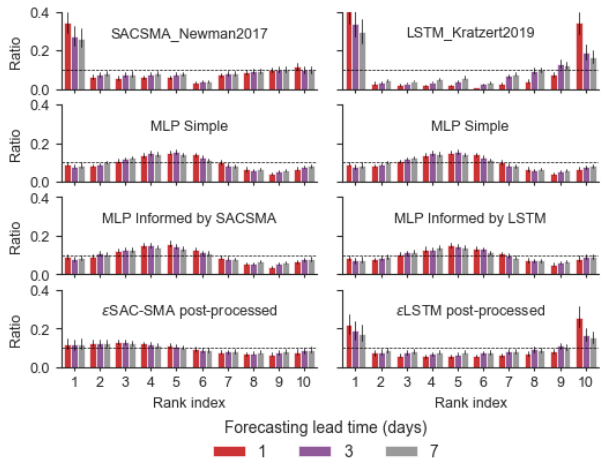

Fig. 2: Rank diagram: benchmark cases in columns, approaches in row.

Following the Fig.2, the findings indicate an atypical L-shape for the SAC-SMA (top left) and a U-shape for the LSTM (top right). For the baseline (row 2), a flat-shape can be noted with a slight dome on mid-ranged flows. When applying the DA techniques (row 2-4), the distribution becomes more reliable, since it looks trying to flatten the prior distributions. The success is higher on the SACSMA cases but remain significant globally.

3.2.2 ROC curve and the AUC (Area under the ROC curve)

The ROC curve and the AUC can be used to assess the accuracy of a forecast to detect events. Perfect (or low quality) forecast models will have $AUC = 1$ (or $AUC \leq 0.5$) respectively. The AUC scores for all cases are shown in Fig. 3. We experiment anticipating floods events on various discharge thresholds for both drought and flood contexts using the evaluation period. The models are

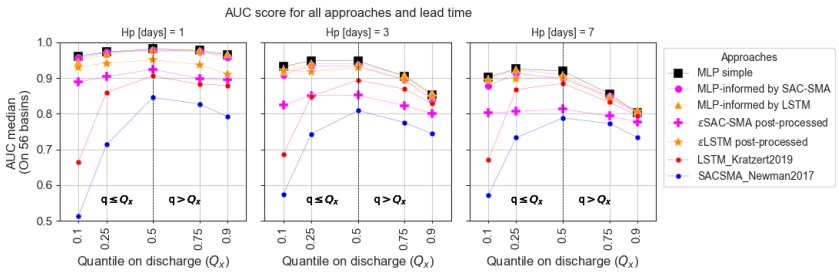

Fig. 3: AUC Scores

quite accurate to detect events in both drought and flood contexts. They look better on floods, and wile the overall performance decreases as the lead time get longer. Unsurprisingly, the benchmark models (without DA) issue the lowest performances, while the baseline which strongly assimilates the discharge data look providing the best. They can globally be ranked as : (1) MLP Simple and MLP Informed by benchmarks; (2) **ePP** cases; (3) LSTM and (4) SACSMA.

4 Conclusion

This study evaluated the performances of two popular hydrologic models (LSTM of [1] and the SACSMA of [2]) under an operational discharge forecast use. The trade-off of the absence of the DA and the persistence analysis in their design were addressed. Using three DA techniques, these benchmarks work have been orchestrated into a forecasting module. The persistence on both the benchmark models and the tested approaches confirms the gap that may exist between simulation and forecasting. The model's sensitivity to the weather uncertainties is strongly improved by the assimilated data, either directly or indirectly. These findings remain relevant to what is usually stated in hydrologic modeling, where, it is harder to forecast on short than long-term lead times; DA is more informative on short-term lead time, DL models outperform traditional models, ... However, further experiments may be engaged to assess the limit of the DA techniques, explore more sampling strategies of the weather data, extend the criteria list in order to reveal other sides of the approaches.

References

[1] F Kratzert, D Klotz, G Shalev, G Klambauer, S Hochreiter, and G Nearing. Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, 23(12):5089–5110, 2019.

[2] Andrew J Newman, Naoki Mizukami, Martyn P Clark, Andrew W Wood, Bart Nijssen, and Grey Nearing. Benchmarking of a Physically Based Hydrologic Model. *Journal of Hydrometeorology*, 18(8):2215–2225, 2017.

[3] A J Newman, M P Clark, K Sampson, A Wood, L E Hay, A Bock, R J Viger, D Blodgett, L Brekke, J R Arnold, T Hopson, and Q Duan. Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences*, 19(1):209–223, 2015.

[4] Bob E Saint Fleur, Guillaume Artigue, Anne Johannet, and SÃ©verin Pistre. Deep Multilayer Perceptron for Knowledge Extraction: Understanding the Gardon de Mialet Flash Floods Modeling. In Olga Valenzuela, Fernando Rojas, Luis Javier Herrera, HÃ©ctor Pomares, and Ignacio Rojas, editors, *Theory and Applications of Time Series Analysis*, pages 333–348, Cham, 2020. Springer International Publishing.

[5] The Data Assimilation Research Testbed (Version X.Y.Z) [Software]. (2024). Boulder, Colorado: UCAR/NSF NCAR/CISL/DAReS., 2024.

[6] Peter K Kitanidis and Rafael L Bras. Real-time forecasting with a conceptual hydrologic model: 2. Applications and results. *Water Resources Research*, 16(6):1034–1044, 1980.

[7] Jens Christian Refsgaard and Jesper Knudsen. Operational Validation and Intercomparison of Different Types of Hydrological Models. *Water Resources Research*, 32(7):2189–2202, 7 1996.

[8] Anna Pölz, Alfred Paul Blaschke, JÃ¼rgen Komma, Andreas H Farnleitner, and Julia Derx. Transformer Versus LSTM: A Comparison of Deep Learning Models for Karst Spring Discharge Forecasting. *Water Resources Research*, 60(4):e2022WR032602, 4 2024.

[9] O Talagrand, R Vautard, and B Strauss. *Evaluation of probabilistic prediction systems*. PhD thesis, Shinfield Park, Reading, 5 1997.

[10] Ian Mason. A model for assessment of weather forecasts. *Aust. Meteor. Mag*, 30(4):291–303, 1982.

[11] W Wesley Peterson, Theodore G Birdsall, and William C Fox. The theory of signal detectability. *Trans. IRE Prof. Group Inf. Theory*, 4:171–212, 1954.