Response

Anonymous Referee #1:

In this paper, four experiments were conducted to assess the performance of three types of Rainfall-Runoff models: Long Short-Term Memory (LSTM), mass conservative LSTM (MC-LSTM), and the conceptual model EXP-HYDRO. The document is well-organized, and the results are presented and discussed clearly.

We are grateful to you for the positive comments.

Nevertheless, certain critical aspects arise in the formulation and execution of the study:

Thank you very much for the constructive comments. We have improved the paper accordingly and provide point-by-point responses.

1. *The primary justification for the study is based on the authors' assertion that "there is yet no consensus on the effects of the water balance constraint on the use of the LSTM network." To support this claim, the authors cite the works of Cai et al. (2022) and Wang et al. (2023), stating higher accuracies with MC-LSTM over LSTM. However, Cai et al. (2022) does not use LSTM neural networks, and Wang et al. (2023) applies physically informed LSTMs for a different purpose and at a scale unrelated to the Rainfall-Runoff models of Frame et al. (2022) and Frame et al. (2023). Consequently, the comparison appears unfair, involving different processes, evaluated at different scales, and with significantly different instances.*

Thank you for the insightful comment. We are sorry for the confusing information and have modified the expression and citation. With similar scales, same processes and comparable instances, Nearing et al. (2020), Wi and Steinschneider (2024) and Frame et al. (2023, 2022) provide fair comparison for rainfall-runoff prediction in the revision. In the meantime, the focus of this paper is emphasized. The revised text is as follows:

"Using an architecturally mass-conserving variant of the LSTM (MC-LSTM) (Hoedt et al., 2021), it has been found that the water balance constraint can enhance the accuracy and extrapolation ability of the LSTM network for rainfall-runoff prediction (Nearing et al., 2020; Wi and Steinschneider, 2024). On the other hand, it has recently been observed that the water balance constraint can impair the predictive performance under extreme events (Frame et al., 2023, 2022), as well as hard constrained models in multitask hydrological forecasting (Li et al., 2024). Therefore, there is yet no consensus on the effects of the water balance constraint on the use of the LSTM network for rainfall-runoff prediction (Pokharel et al., 2023), particularly its robustness—the ability to perform consistently across varying conditions (Manure et al., 2023)" (Page 3, Lines 55 to 62)

2. *Considering the preceding point, the objective and main conclusion of the work are not clear. While each of the four experiments is thoroughly described, there appears to be a lack of novelty. For instance, the results of experiments 1 and 2 are somewhat expected and have already been published by other authors using diverse datasets.*

Thank you for the constructive comments. We have rewritten the Discussion section to emphasize the significance of this paper:

"It has been highlighted that more training data contributes to the performance of the LSTM network (Gauch et al., 2021b; Read et al., 2019; Wang et al., 2023; Xie et al., 2021). In the case of different training data amounts, the role of the water balance constraint in the performance of the LSTM network is investigated through large-sample tests in different aspects. For single-basin trained DL and TGDS models, previous studies have quantified how additional training data improve their predictions in limited areas (Read et al., 2019; Wang et al., 2023), rather than large-sample tests which can help to understand model limitations and draw robust conclusions from a big-picture perspective (Addor et al., 2020; Gupta et al., 2014). In addition, a modified

DSST method is utilized to assess the transferability of the MC-LSTM and LSTM under contrasting climate conditions in this paper. Recent studies have assessed the credibility of future streamflow projections under warming through metamorphic testing (Razavi, 2021; Reichert et al., 2023; Wi and Steinschneider, 2022, 2024; Yang and Chui, 2021), while metamorphic testing requires partly subjective expert judgements and utilizes no realistic climate change scenarios (Reichert et al., 2023), thereby may leading to unreliable results (Wi and Steinschneider, 2024).” (Pages 18 to 19, Lines 364 to 374)

3. *As highlighted by Kratzert et al. (2024, https://doi.org/10.5194/hess-2023-275), the training of effective rainfall-runoff LSTM models requires the use of multiple basins. However, in this study, several LSTM models are trained with single-basin data. Consequently, suboptimal LSTM models are likely obtained, and the superior performance demonstrated by MC-LSTM may be a consequence of these suboptimal models.*

Thank you for the insightful comments. To clarify the differences and connections between this paper with the mentioned literature above, we have rewritten the Discussion section and the related part are as follows:

“Although TGDS models can provide more accurate and robust predictions than pure DL models in basin-wise scale or data scarce conditions, it deserves additional scrutiny when trained with data from a large number of diverse basins (Frame et al., 2022; Nearing et al., 2021; Wi and Steinschneider, 2022). Recent studies have illustrated that the LSTM network works better for rainfall-runoff prediction when trained with a large amount of hydrologically diverse data than with data from a single watershed (Kratzert et al., 2024). Specifically, for DL models, physical constraints are effective in local models but offer little improvement in the regional models (Frame et al., 2023; Xie et al., 2021), even reduce predictive performance under extreme events (Frame et al., 2022). This outcome can be attributed to that pure DL models might be flexible enough to capture the behaviour in observation data with inconsistent water balance closure

better than DL models constrained by the strict water balance (Kratzert et al., 2024; Frame et al., 2023; Beven, 2020). Besides, catchments with similar flood generating processes and similar characteristics may have some similar outliers and DL models can capture the rainfall-runoff responses among these basins (Xie et al., 2021; Bertola et al., 2023; Wi and Steinschneider, 2024). Therefore, there seems to be a compensating effect between data and knowledge on DL models, where the process knowledge is crucial for models trained with sparse data but less important with sufficient data. Large-sample hydrology is thus expected to enhance the performances of DL models for extreme events predictions and climate change projections (Bertola et al., 2023; Wi and Steinschneider, 2022, 2024).

Given that data is not always sufficient, the sensitivity of DL models when given scarce training data is essentially important (Feng et al., 2021; Gauch et al., 2021b). The TGDS provides effective tools for reducing data requirements of DL models (Karniadakis et al., 2021; Karpatne et al., 2017; Read et al., 2019; Xie et al., 2021). Therefore, this paper explores the effects of the water balance constraint on the robustness of the LSTM under restricted conditions, thereby training single model for each single basin rather than simultaneously for a large number of basins. Although the latter can achieve better performance (Kratzert et al., 2024), it is beyond the scope of this paper but worthy of further study…" (Page 19, Lines 378 to 397)

4. *It is unclear why the authors opted for the hyperparameters listed in Table 1. Moreover, it is essential to consider how the results of their experiments might be influenced by alternative hyperparameters. Could the enhanced performance of MC-LSTM be subject to change with different values of model hyperparameters?*

Thank you for the constructive comments. We are sorry for the unclear information and have added the tunning processes in the Supplement as Text S2:

"There are two categories of hyperparameters to be tuned, including hyperparameters for model structure (i.e., hidden layer, hidden size) and hyperparameters for the training

process (i.e., learning rate, batch size) (Li et al., 2024). Given a balance of model performance and time cost, 50 basins are selected randomly for the tuning processes. For each catchment, the first 14 years (from 1 October 1980 to 30 September 1994) of the entire training period (from 1 October 1980 to 30 September 1995) is set as the training period for tuning process and the last year (from 1 October 1994 to 30 September 1995) is set as the validation period. Models are trained using the Adam optimizer and the early stopping strategy. Three repetitions of each hyperparameter setting are used with different random seeds for initializing the weights, the mean NSE on validation period over the 3 repetitions represents the validation performance of a basin. Hyperparameters were chosen using the model settings with the highest median NSE scores on validation period over the 50 basins.

Firstly, the model structure hyperparameters are fine-tuned. Based on the model structure of Kratzert et al. (2018) (two hidden layers with the hidden size of 20) and the results that a one-layer LSTM network is qualified to capture rainfall-runoff response of a catchment (Kratzert et al., 2019, 2021), the range of the hidden size in this paper is set to 20，40，50，60, 80 and 100 while the number of hidden layers is set to 1. Following other hyperparameters of Kratzert et al. (2018), the LSTM is developed with input sequence for the past T = 365 d, a mini-batch size of 512, a drop-out rate of 0.1 and the Adam optimizer with a learning rate of 0.0001. Secondly, the hyperparameters for the training process is optimized based on the optimal hyperparameters in the first step. The LSTM network is tuned with different batch sizes (128, 256, 512), different learning rates (0.1, 0.01, 0.001, 0.0001), different learning rate decay (0.1, 0.3, 0.5, 0.7) and different dropout rates (0.2, 0.4, 0.6, 0.8).

After tunning, the optimal hyperparameters of the LSTM are shown by Table 1. In order to compromise between maximum reducing the uncertainty caused by different numbers of model parameters and achieving potentially more powerful predictions, the hidden sizes of the MC-LSTM network is set to 50, respectively, so that the numbers of parameters for MC-LSTM and LSTM differ by less than 0.1%. As the EXP-HYDRO model is a process-based model, there is no need for the DL wrapped EXP-HYDRO

model to normalize their input variables and to set the hidden size or dropout rate. Excluding the hidden size and dropout rate, the MC-LSTM and EXP-HYDRO models in the four experiments have the same hyperparameters as the LSTM, as shown by Table 1. Notably, the MC-LSTM has some hyperparameters from the LSTM instead of being optimized, while tunning the hyperparameters of the MC-LSTM can obtain better MC-LSTM networks. However, this paper aims to investigate the robustness of the LSTM and MC-LSTM rather than thoroughly explore the potential of the water balance constraint on the use of the LSTM network for rainfall-runoff prediction. Thus, further tuning processes of the MC-LSTM are not performed. Furthermore, the sensitivity analysis of model hyperparameters are devised based on model hyperparameters in Frame et al. (2023, 2022), so the hidden sizes of the LSTM and MC-LSTM are 256 and 64, respectively. The results of the sensitivity analysis of model hyperparameters are presented by Fig. S3 to S8 in the Supplement."

In addition, to consider the results of the experiments in this paper might be influenced by model hyperparameters, we have devised an additional experiment to demonstrate the reliability of the results in the Discussion section:

"Besides, the results of sensitivity analysis of model hyperparameters are presented by Fig. S3 to S8 in the Supplement. It can be observed that the enhanced robustness of the MC-LSTM compared with the LSTM changes little with different model hyperparameters, which demonstrates the reliability of the results in this paper." (Page 19, Lines 374 to 377)

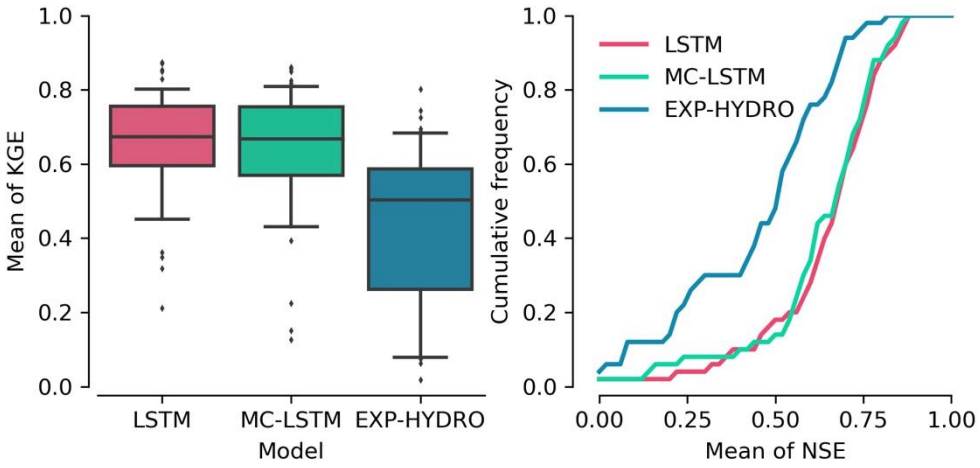

Figure S3. As for Fig. 2, but for the LSTM and MC-LSTM with hidden sizes of 256 and 64, respectively, in 50 randomly selected basins.

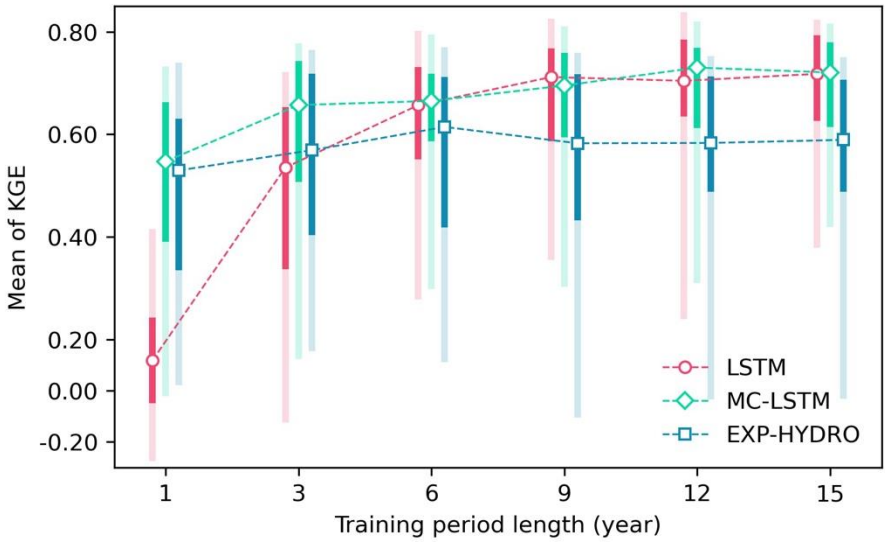

Figure S4. As for Fig. 4, but for the LSTM and MC-LSTM with hidden sizes of 256 and 64, respectively, in 50 randomly selected basins.

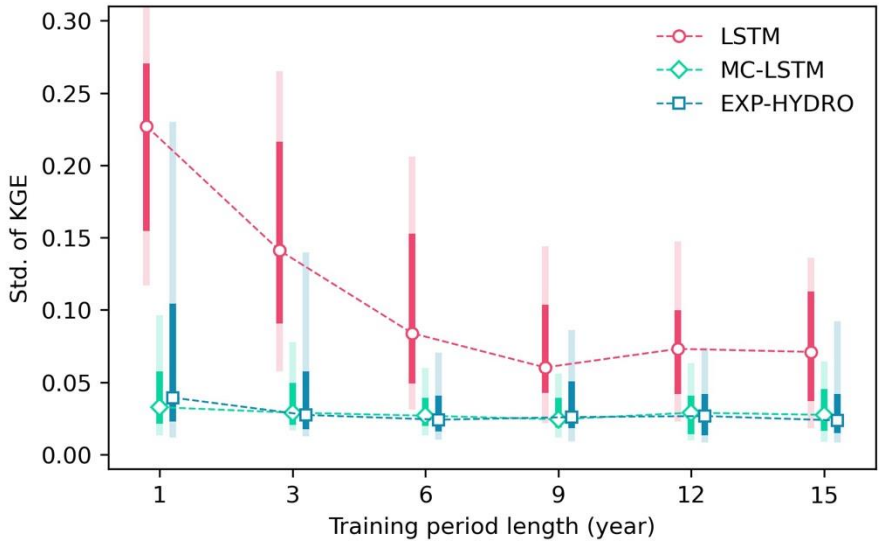

Figure S5. As for Fig. 5, but for the LSTM and MC-LSTM with hidden sizes of 256 and 64, respectively, in 50 randomly selected basins.

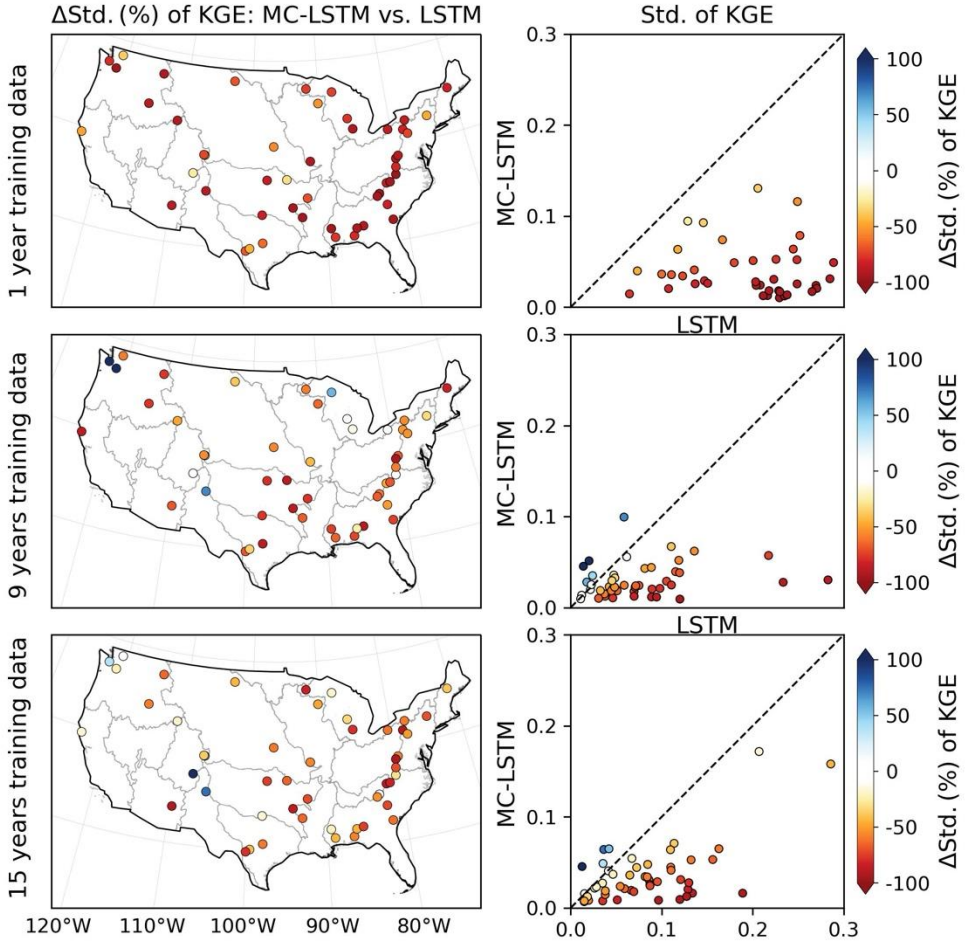

Figure S6. As for Fig. 6, but for the LSTM and MC-LSTM with hidden sizes of 256 and 64, respectively, in 50 randomly selected basins. The MC-LSTM tends to be more stable at a total of 50 (100%), 43 (86%) and 45 (90%) basins when models are trained with data of 3 years, 9 years and 15 years, respectively.

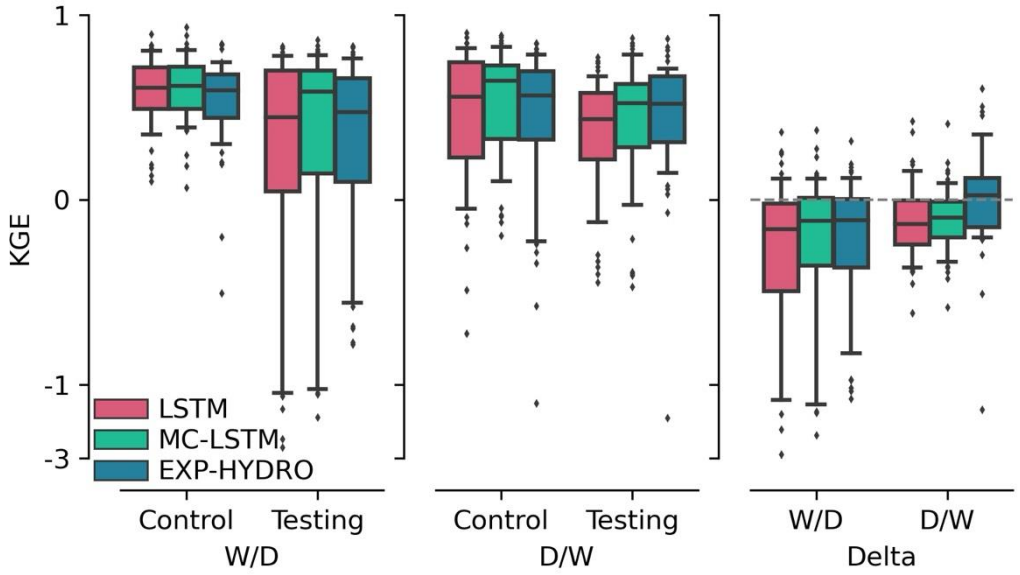

Figure S7. As for Fig. 7, but for the LSTM and MC-LSTM with hidden sizes of 256 and 64, respectively, in 50 randomly selected basins.

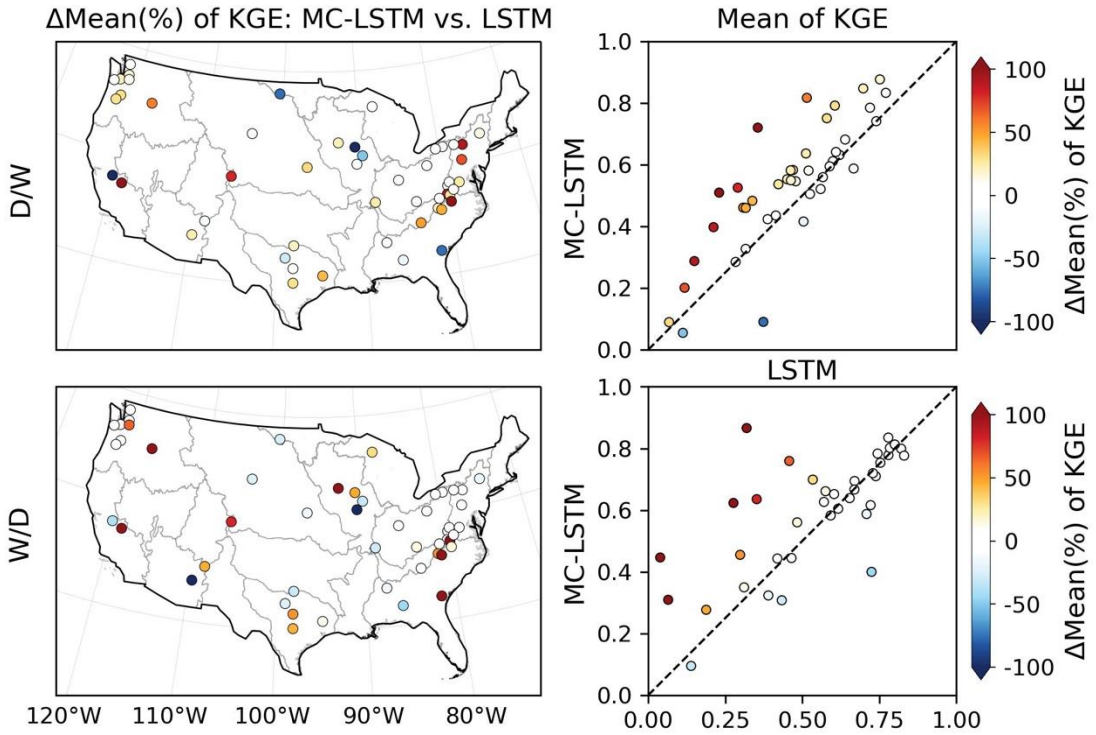

Figure S8. As for Fig. 8, but for the LSTM and MC-LSTM with hidden sizes of 256 and 64, in 50 randomly selected basins. In the D/W scenario, the MC-LSTM exhibits higher KGE values compared to the LSTM across 38 basins (76%). But for the W/D scenario, the number of basins with higher KGE for the MC-LSTM than the LSTM decreases to 26 (52%).

References:

Frame, J. M., Kratzert, F., Klotz, D., Gauch, M., Shalev, G., Gilon, O., Qualls, L. M., Gupta, H. V., and Nearing, G. S.: Deep learning rainfall–runoff predictions of extreme events, Hydrol. Earth Syst. Sci., 26, 3377–3392, https://doi.org/10.5194/hess-26-3377-2022, 2022.

Frame, J. M., Kratzert, F., Gupta, H. V., Ullrich, P., and Nearing, G. S.: On strictly enforced mass conservation constraints for modelling the rainfall-runoff process, Hydrol. Process., 37, e14847, https://doi.org/10.1002/hyp.14847, 2023.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M.: Rainfall-runoff modelling using long short-term memory (LSTM) networks, Hydrol. Earth Syst. Sci., 22, 6005–6022, https://doi.org/10.5194/hess-22-6005-2018, 2018.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., and Nearing,

G. S.: Toward improved predictions in ungauged basins: exploiting the power of machine learning, Water Resour. Res., 55, 11344–11354, https://doi.org/10.1029/2019WR026065, 2019.

Kratzert, F., Klotz, D., Hochreiter, S., and Nearing, G. S.: A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall-runoff modeling, Hydrol. Earth Syst. Sci., 25, 2685–2703, https://doi.org/10.5194/hess-25-2685-2021, 2021.

Li, L., Dai, Y., Wei, Z., Shangguan, W., Zhang, Y., Wei, N., and Li, Q.: Enforcing Water Balance in Multitask Deep Learning Models for Hydrological Forecasting, J. Hydrometeorol., 25, 89–103, https://doi.org/10.1175/JHM-D-23-0073.1, 2024.

Nearing, G., Kratzert, F., Klotz, D., Hoedt, P.-J., Klambauer, G., Hochreiter, S., and Gupta, H.: A deep learning architecture for conservative dynamical systems: application to rainfall-runoff modeling, in: AI for Earth Sciences Workshop, NeurIPS 2020, 2020.

Wi, S. and Steinschneider, S.: On the need for physical constraints in deep learning rainfall–runoff projections under climate change: a sensitivity analysis to warming and shifts in potential evapotranspiration, Hydrol. Earth Syst. Sci., 28, 479–503, https://doi.org/10.5194/hess-28-479-2024, 2024.