Hydrology Research

Hydrology Research Vol 53 No 5, 657 doi: 10.2166/nh.2022.044

Evaluating the long short-term memory (LSTM) network for discharge prediction under changing climate conditions

Carolina Natel de Moura [a,*], Jan Seibert[b] and Daniel Henrique Marco Detzel[a]

[a] Department of Hydraulics and Sanitation, Federal University of Parana, Avenida Coronel Francisco Heráclito dos Santos, 100, 81531-980 Curitiba, Paraná, Brazil
[b] Department of Geography, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland
*Corresponding author. E-mail: carolina.natel@ufpr.br

CN, 0000-0003-3103-6789

ABSTRACT

Better understanding the predictive capabilities of hydrological models under contrasting climate conditions will enable more robust decision-making. Here, we tested the ability of the long short-term memory (LSTM) for daily discharge prediction under changing conditions using six snow-influenced catchments in Switzerland. We benchmarked the LSTM using the Hydrologiska Byråns Vattenbalansavdelning (HBV) bucket-type model with two parameterizations. We compared the model performance under changing conditions against constant conditions and tested the impact of the time-series size used in calibration on the model performance. When calibrated, the LSTM resulted in a much better fit than the HBV. However, in validation, the performance of the LSTM dropped considerably, and the fit was as good or poorer than the HBV performance in validation. Using longer time series in calibration improved the robustness of the LSTM, whereas HBV needed fewer data to ensure a robust parameterization. When using the maximum number of years in calibration, the LSTM was considered robust to simulate discharges in a drier period than the one used in calibration. Overall, the HBV was found to be less sensitive for applications under contrasted climates than the data-driven model. However, other LSTM modeling setups might be able to improve the transferability between different conditions.

Key words: climate transposability, data-driven model, differential split-sample test, model calibration, model robustness

HIGHLIGHTS

- The long short-term memory (LSTM) had good predictive accuracy in both the calibration and validation periods; however, it is always less robust than the HBV model.
- When using the maximum number of years in calibration, the LSTM was robust enough in its application under changing conditions when applied in a condition drier than the one used in calibration.

INTRODUCTION

The use of hydrological models in conditions that differ from those during model calibration is a challenging problem in hydrology and critical for application in impact studies (Blöschl *et al.* 2019). Models calibrated in certain conditions have been shown to be not always suitable for different conditions or transferable in time (Bastola *et al.* 2011; Coron *et al.* 2012; Thirel *et al.* 2015; Broderick *et al.* 2016; Dakhlaoui *et al.* 2017; Grusson *et al.* 2017; Her *et al.* 2019; Ouermi *et al.* 2019; Pan *et al.* 2019). The lack of a robust analysis of model performance under changing conditions may lead to poor water resource management.

In the context of catchment hydrology, a changing condition refers to any significant modification in land cover, climate, or water management infrastructure, potentially affecting the transformation of rainfall into runoff (Thirel *et al.* 2015). A general approach for developing hydrological models suitable for use in transient conditions is to use the differential split-sample test (DSST). The model should be calibrated and validated over contrasting periods in such a method, for instance, calibrated over a wet period and validated during a dry period (Klemeš 1986; Coron *et al.* 2012). The modeler should seek a good transferability of the calibrated parameters to a different dataset in validation, rather than only a good fit during calibration, which is often translated as model robustness. Robustness is a model's degree of insensitivity to climatic and environmental conditions (Seiller *et al.* 2012).

Model generalization for contrasting climates has been extensively explored in the literature using the DSST (Seibert 2003; Wilby 2005; Vaze *et al.* 2010; Merz *et al.* 2011; Coron *et al.* 2012; Li *et al.* 2012; Seiller *et al.* 2012; Brigode *et al.* 2013; Kling *et al.* 2015; Li *et al.* 2015; Seiller *et al.* 2015; Thirel *et al.* 2015; Broderick *et al.* 2016; Fowler *et al.* 2016; Vormoor *et al.* 2018). The results have shown that model parameters are sensitive to the climatic conditions of the calibration period (Pan *et al.* 2019), that the transfer of model parameters in time may introduce a significant level of simulation errors (Zhu *et al.* 2016), and that calibration over a wetter (drier) climate than the validation climate leads to an overestimation (underestimation) of the mean simulated runoff (Coron *et al.* 2012). Changes in mean rainfall were more likely than those in mean potential evapotranspiration or air temperature to impact performance during validation (Coron *et al.* 2012). Furthermore, Broderick *et al.* (2016) pointed out that the model transferability in contrasted climates may vary depending on the testing scenario, catchment, and evaluation criteria. Here, we argue that testing new models and new calibration protocols can help with our understanding of the modeling capabilities under changing conditions.

Although data-driven techniques have proved to outperform many traditional approaches based on conceptual or physical models for constant conditions (Dawson & Wilby 1998; Dibike & Solomatine 2001; Hu *et al.* 2018; Lee *et al.* 2018; Kratzert *et al.* 2019a; Rafaeli Neto *et al.* 2019; Xu *et al.* 2020), and the models are reliable in out-of-sample generalization (Shen 2018), little work has been carried out to test the capabilities of data-driven methods to make reasonable predictions under changing conditions. A significant limitation of data-driven models may be that they do not benefit from our understanding of physical phenomena and instead rely on the data provided during optimization. Shortridge *et al.* (2016) argued that data-driven models could only generate reliable predictions for conditions comparable to those experienced historically. Otherwise, the models are likely to introduce considerable uncertainty into their projections.

The long short-term memory (LSTM), a particular type of recurrent neural network (RNN), has been shown to be promising in capturing the hydrological behavior from the learning process (Xu *et al.* 2020). Lees *et al.* (2021) showed that the LSTM simulates discharge with a consistent high model performance in a large range of catchments in Great Britain, including catchments typically considered difficult to model with four lumped conceptual models. Kratzert *et al.* (2019b) applied the LSTM model over 531 basins over the USA and found a high correlation between the values of the internal cells of an LSTM network and natural processes.

Recently, O & Orth (2020) evaluated state-of-the-art models to changing conditions, calibrating an LSTM network and two process-based models in 161 catchments distributed across Europe. In their modeling setup, the LSTM model and the process-based models had different calibration approaches. The LSTM was calibrated over all catchments at once using two approaches: calibrating on an extreme reference period (365 days) and calibrating with one randomly selected year from each catchment rather than the respective extreme reference year. In contrast, the process-based models were calibrated in individual catchments and only using the extreme reference period. The models were then used to simulate in the remaining years characterized by a transient condition. The models showed overall performance loss, which generally increased the more conditions deviated from the reference climate, and overall, relatively high robustness was demonstrated by the physically-based model.

In light of the discussion above, in this paper we tested new calibration protocols and extended the scope of the model evaluation, with focus on the LSTM model. This is done by (a) benchmarking the LSTM using the same modeling setup for both data-driven and process-based models (which includes calibrating one model to each catchment instead of calibrating the LSTM over all catchments), (b) testing if increasing the number of years in model calibration would lead to better model performance and robustness, in contrast to only 1 year used in the previous study), and finally (c) calibrating the models in constant conditions as comparison. We then evaluated the robustness of the LSTM for contrasted conditions compared to both its application in constant conditions and the robustness obtained by the process-based model.

STUDY AREA AND DATA

For our study, we used six snow-influenced catchments located in Switzerland, ranging from ~60 to 400 km^2, with a mean altitude between ~500 and 1,200 m.a.s.l. The location and description of the catchments (location, area, altitude, daily mean temperature, annual precipitation, mean daily discharge, and snow fraction) are presented in Figure 1 and Table 1. Our catchment choice aimed to select catchments mainly located in the Swiss plateau, within a climate homogeneous area, and considered nearly natural (i.e., there is negligible impact on runoff from human activity) (Orth *et al.* 2015).

The data needed to model the daily discharge were air temperature (°C), precipitation (mm d^{-1}), and the estimates of long-term monthly potential evapotranspiration (mm month^{-1}). Precipitation and air temperature data were obtained from the

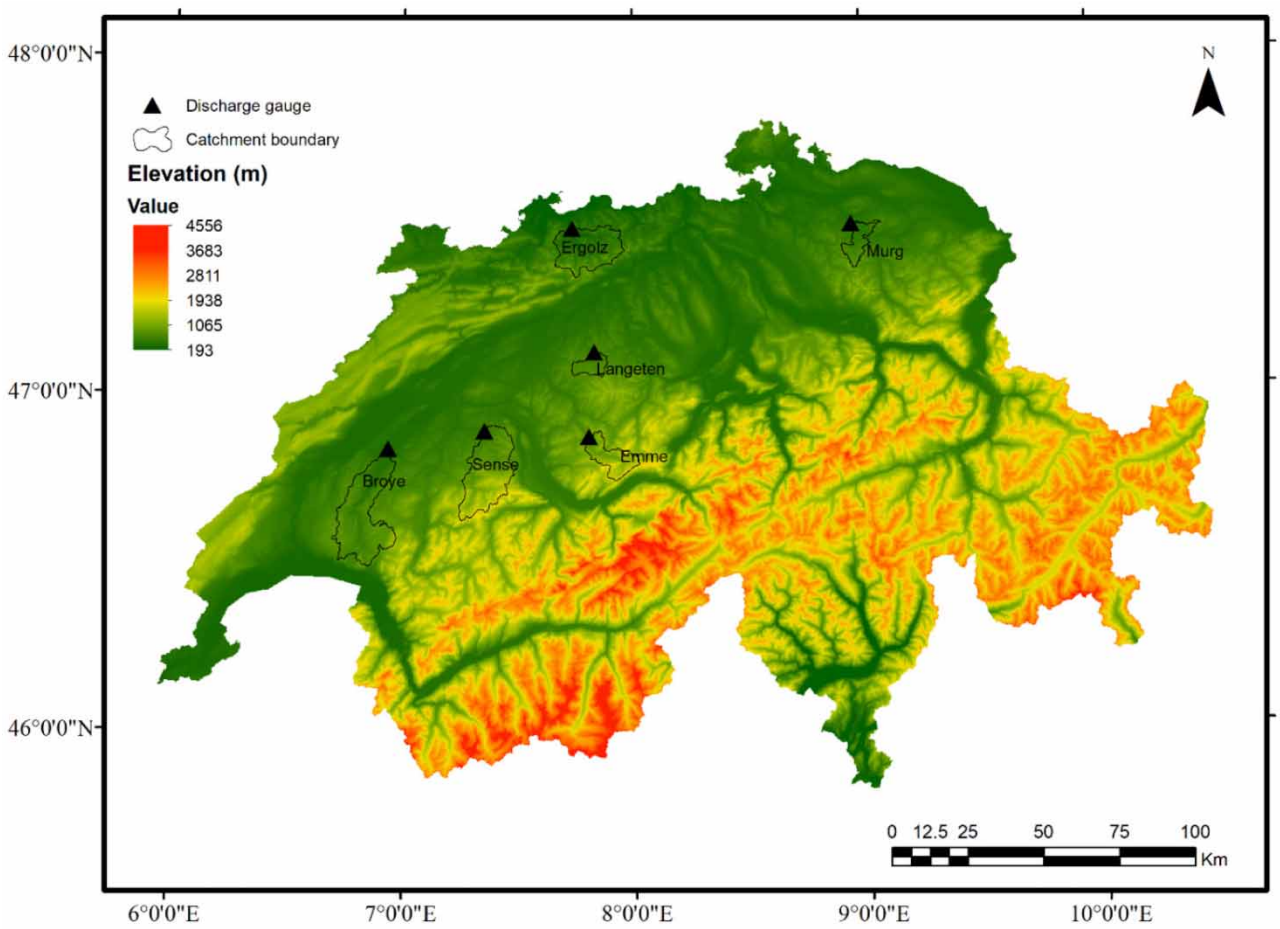

Figure 1 | Location of the study catchments.

Table 1 | Properties of the study catchments

Catchment	Mean altitude (m)	Area (km^2)	Daily mean temperature (°C)	Total precipitation (mm year^{-1})	Mean discharge (mm d^{-1})	Snow fraction[a] (%)
Broye	710	392	8.6	1,190	1.6	5
Emme	1,189	124	5.6	1,692	3.0	19
Ergolz	590	261	8.6	1,091	1.2	6
Langeten	766	60	7.5	1,305	1.8	10
Murg	650	79	8.0	1,313	2.0	7
Sense	1,068	352	6.3	1,445	2.1	13

[a]Snow fraction (%): fraction of precipitation falling with temperature below 0 °C.

gridded meteorological forcing data at the spatial resolution of 2 km × 2 km from the Swiss Federal Office of Meteorology and Climatology (MeteoSwiss). We obtained daily discharge measurements from the Swiss Federal Office for the Environment (FOEN).

METHODS

Long short-term memory (LSTM)

The LSTM is a particular type of RNN used to process long time-sequences of data (Hochreiter & Schmidhuber 1997), in which the output of each time-step is fed as input to the next time-step. The control of the information flow is managed in

units called gates and memory cells. The cell remembers values over arbitrary time intervals, and three gates regulate the flow of information into and out of the cell: the forget gate, the input gate, and the output gate. At every time-step t, each of the three gates is presented with the input $x[t]$ (i.e., explanatory variables) as well as the output $h[t-1]$ of the memory cells at the previous time-step $[t-1]$.

The first gate is the forget gate, which controls what information is removed from the cell state vector (Equation (1)). The hidden state h is initialized in the first time-step by a vector of zeros. In the next step, a potential update vector for the cell state is computed from the current input $x[t]$ and the last hidden state $h[t-1]$ by Equation (2). Additionally, the second gate is computed, the input gate (Equation (3)), defining which (and to what degree) information of $\bar{c}[t]$ is used to update the cell state in the current time-step. With the results of the forget gate and input gate, the cell state $c[t]$ is updated by Equation (4). Like the hidden state vector, the cell state is initialized by a vector of zeros in the first time-step. The last gate is the output gate, which controls the information of the cell state $c[t]$ that flows into the new hidden state $h[t]$. The output gate is calculated by Equation (5). Finally, the hidden state $h[t]$ is calculated by using the current cell state and the output gate value (Equation (6)). The model output is a linear combination of hidden states at the last time-step (Kratzert *et al.* 2018).

$$f[t] = \sigma(W_f x[t] + U_f h[t-1] + b_f) \tag{1}$$

$$\bar{c}[t] = \tan h(W_g x[t] + U_g h[t-1] + b_g) \tag{2}$$

$$i[t] = \sigma(W_i x[t] + U_i h[t-1] + b_i) \tag{3}$$

$$c[t] = f[t] \otimes c[t-1] + i[t] \otimes \bar{c}[t] \tag{4}$$

$$o[t] = \sigma(W_o x[t] + U_o h[t-1] + b_o) \tag{5}$$

$$h[t] = o[t] \otimes \tan h\,(c[t]) \tag{6}$$

where $f[t]$, $i[t]$, and $o[t]$ are the forget, input, and output gates represented by vectors with values in the range (0, 1), $\bar{c}[t]$ is a vector with values in the range (−1, 1), $x[t]$ is the input vector (forcings and static attributes), $\tan h\,(\cdot)$ is the hyperbolic tangent, $\sigma\,(\cdot)$ represents the logistic sigmoid function, $\otimes$ denotes element-wise multiplication, and Ws, Us, and bs are sets of learnable parameters (i.e., two adjustable weight matrices and a bias vector).

In this work, we used a network consisting of a single LSTM layer with one hidden unit and a dense layer that connects the output of the LSTM at the last time-step to a single output neuron with linear activation. The LSTM model was implemented using the Keras package in Python, the Adam activation function, and the mean-squared error as loss function. To predict the discharge of a single time-step (day), we provided as input the last t consecutive time-steps of independent meteorological variables (daily precipitation [mm d^{-1}] and air temperature [°C]). We obtained the best hyperparameters of the LSTM model through a trial-and-error tuning approach. We varied the values of the following hyperparameters: length of the input sequence (time-steps), number of neurons in the hidden layer, and number of epochs. Our analysis resulted in the selection of 50 neurons, 50 epochs, and 365 days as time-steps.

HBV model

We benchmarked the performance of the LSTM model against the bucket-type HBV-Light version model (Seibert & Vis 2012). The HBV model consists of four routines including the snow routine, the soil routine, the groundwater routine, and the routing routine. This model usually simulates daily discharge based on daily precipitation, daily air temperature, and estimates of long-term monthly potential evapotranspiration rates. The HBV was used as both a lower and upper benchmark with two different parameterization methods (Seibert *et al.* 2018). As a lower benchmark, we used the ensemble mean of simulations with 1,000 randomly selected parameter sets, referred to hereafter as 'uncalibrated HBV'. For the upper benchmark, we calibrated the HBV model using an automatic genetic algorithm and the Nash–Sutcliff efficiency (NSE) as objective function, referred to hereafter as 'calibrated HBV'. In both cases, we specified feasible parameter ranges based on previous model applications.

Calibration procedure

The LSTM and HBV were calibrated individually for each one of the catchments resulting in six LSTM models and six HBV models. We calibrated and validated the models according to the DSST proposed by Klemeš (1986) for changing conditions.

According to Klemeš (1986), if the model is intended to simulate streamflow in a wet climate scenario, then it should be calibrated on a dry period of the historical record and validated on a wet period and *vice versa*. Additionally, we calibrated and validated a model under constant conditions.

Selection of the calibration and validation periods

The period between 1961 and 2018 was used to select the constant and changing condition periods. We mimicked the changing conditions by selecting two continuous periods in the time series with different hydrological conditions in the historical record. The dry and wet periods were chosen as the annual discharge below and above the long-term average discharge, respectively. The discharge changes between the periods were on average 50%. This is similar to the future hydrological changes expected for Switzerland of an increase in mean and maximum floods of 5–24% in the near future and of 25–49% in the far future, with exception to the Southern alpine catchments, where the mean annual floods may decrease in the far future (Köplin *et al.* 2014). For the constant conditions, we selected continuous periods containing both dry and wet years.

We also selected calibration periods with different time-series sizes, ranging from 2 to 6 years (2, 3, 4, and 6 years) for each catchment and condition (constant and changing), to test the influence of the amount of data used in the calibration on the model performance. We limited this analysis to 6 years due to data availability. We needed continuous periods with only low or high discharge, which were limited, on average, to 6 years across all the catchments.

Evaluation metrics and robustness

We evaluated the model performance using the NSE (Nash & Sutcliffe 1970), Kling–Kupta efficiency (KGE) (Gupta *et al.* 2009), non-parametric efficiency (NPE) (Pool *et al.* 2018), and mean absolute relative error (MARE) (Staudinger *et al.* 2011). The metrics range from $-\infty$ to 1, where 1 indicates perfect agreement between simulations and observations, and values lower than zero indicate very poor performance. These metrics were chosen to evaluate different hydrograph phases, the NSE focus on peaks and discharge dynamics, the KGE focus on the mean, variability, and dynamic, the NPE is the non-parametric version of the KGE, and finally, the MARE focus on low-to-medium flows. The robustness was calculated as the difference between the efficiency in calibration and validation (Hallouin *et al.* 2020). The independent two-sample *t*-test was used to evaluate whether the LSTM mean robustness was equal to the robustness obtained with the HBV model, and to compare the mean robustness of the LSTM under changing and constant conditions, at the significance level (α) of 0.05.

RESULTS

Model performance

In the calibration mode, the LSTM performed better than the HBV model for all criteria as expected, since it is more flexible (it has more degrees of freedom) than the conceptual model. However, the performance of the LSTM decreased more than the calibrated HBV when switching to the validation periods (Figure 2). The uncalibrated HBV model performed less well, but the performance was still better than what one might expect from a model run with random parameters and/or no local information. Therefore, we considered that a model performance of about 0.5 for NSE basically indicates that a model has no skill. By definition, its performance did not systematically differ between the calibration and validation periods for the uncalibrated model. For KGE, the patterns were roughly similar, whereas for NPE and MARE, which are more different from the NSE used for calibration, the calibrated models (LSTM and HBV) were less superior compared to the uncalibrated HBV model, especially when using fewer years during calibration.

The effect of the time-series size used in calibration on the performance of the models is represented in the *x*-axis of Figure 2. There was a positive correlation between the time-series length and model performances, which was more pronounced for the LSTM model. When evaluating the model's performance against metrics not used for the optimization of the model (i.e., KGE, NPE, and MARE), the increase in the time-series length used in calibration is essential to obtain LSTM performances comparable to the HBV model during the validation for contrasted conditions. Simulations for changing conditions performed less well than those for constant conditions in validation. However, the differences were less pronounced using the maximum number of years in calibration (i.e., 6 years).

The hydrographs and scatter plots of observed and estimated discharge using the best configuration, that is, using 6 years in calibration, for one of the study catchments are presented in Figures 3 and 4, respectively. The hydrograph shows the

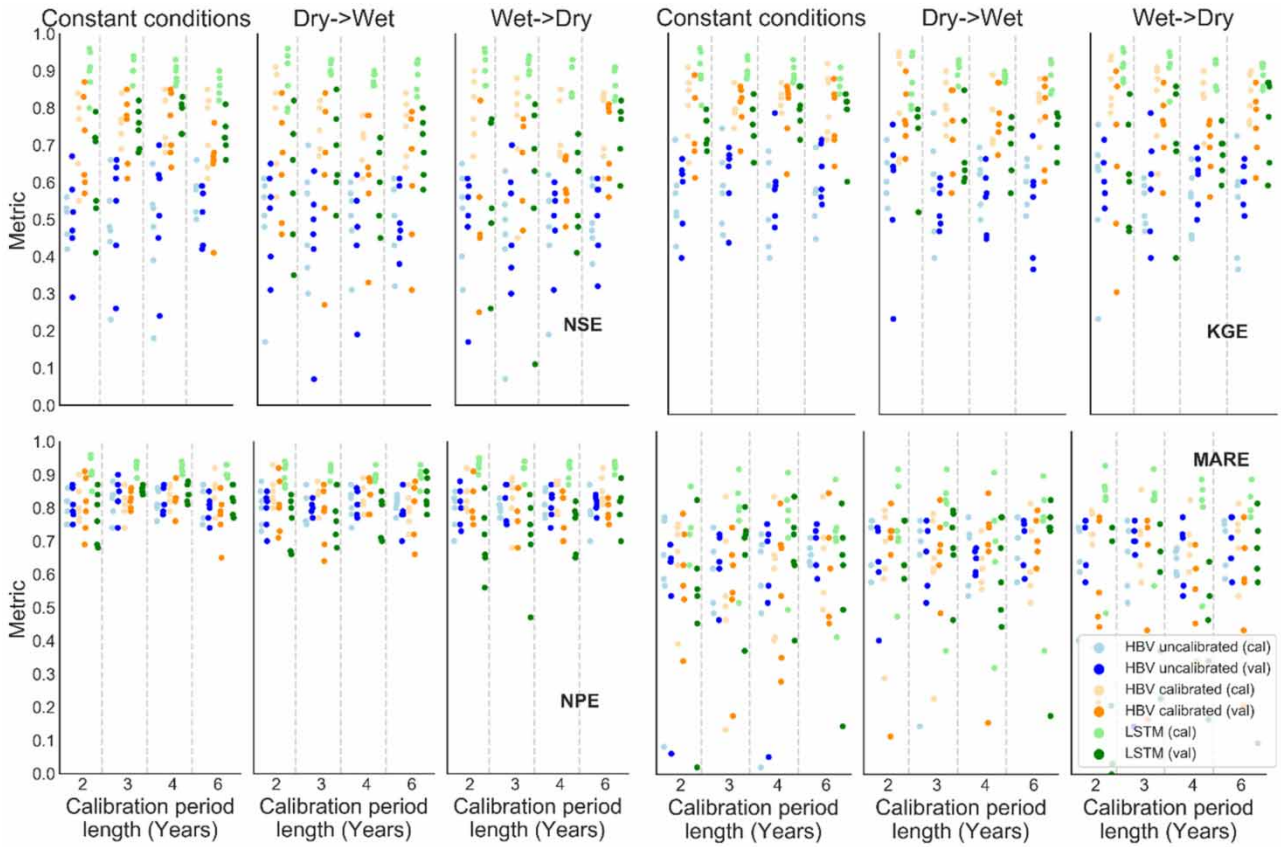

Figure 2 | Dot plots of model performance in calibration (cal) and validation (val) periods for the six catchments under the study. Each grid is one metric, and within each grid, each subgrid is one condition, from left to right: constant conditions, Dry → Wet (calibration in a dry period and validation in a wet one), and Wet → Dry (calibration in a wet period and validation in a dry one). The *y*-axis was limited to the interval between 0 and 1.

underestimation of the peaks, especially those in spring (when the snow accumulated during the winter starts to melt) by all models. However, most low- and mid-flows were predicted well. This is clearly shown in the scatter plots of the observed and simulated flows in Figure 4. The scatter plots also indicate that the predictions deviate more from the observed values in the uncalibrated HBV model. There is an underestimation of the peaks when applying the model in conditions wetter than those it was calibrated in, and the LSTM model simulations are slightly less spread than those of the calibrated HBV model.

Model robustness

The model robustness was evaluated as the difference in performance between calibration and validation periods (Table 2). The LSTM was considered robust enough for generalization in changing conditions when the LSTM mean robustness did not significantly differ from both the mean robustness of the bucket-type model and of the constant period for most of the metrics.

The calibrated HBV was always more robust than the LSTM model for both constant and non-constant conditions. The LSTM was robust enough for changing conditions only when the model was applied in a drier period than that used in calibration and using the maximum number of years during calibration (6 years). While a good indication of robustness was already observed with a shorter time series used in the calibration for the HBV, a longer dataset length was needed for the LSTM.

DISCUSSION

The LSTM had poorer performance under changing conditions. Others have found similar results when applying process-based models under changing conditions (Refsgaard & Knudsen 1996; Xu 1999; Seibert 2003; Wilby 2005; Chiew *et al.* 2009; Vaze *et al.* 2010; Bastola *et al.* 2011).

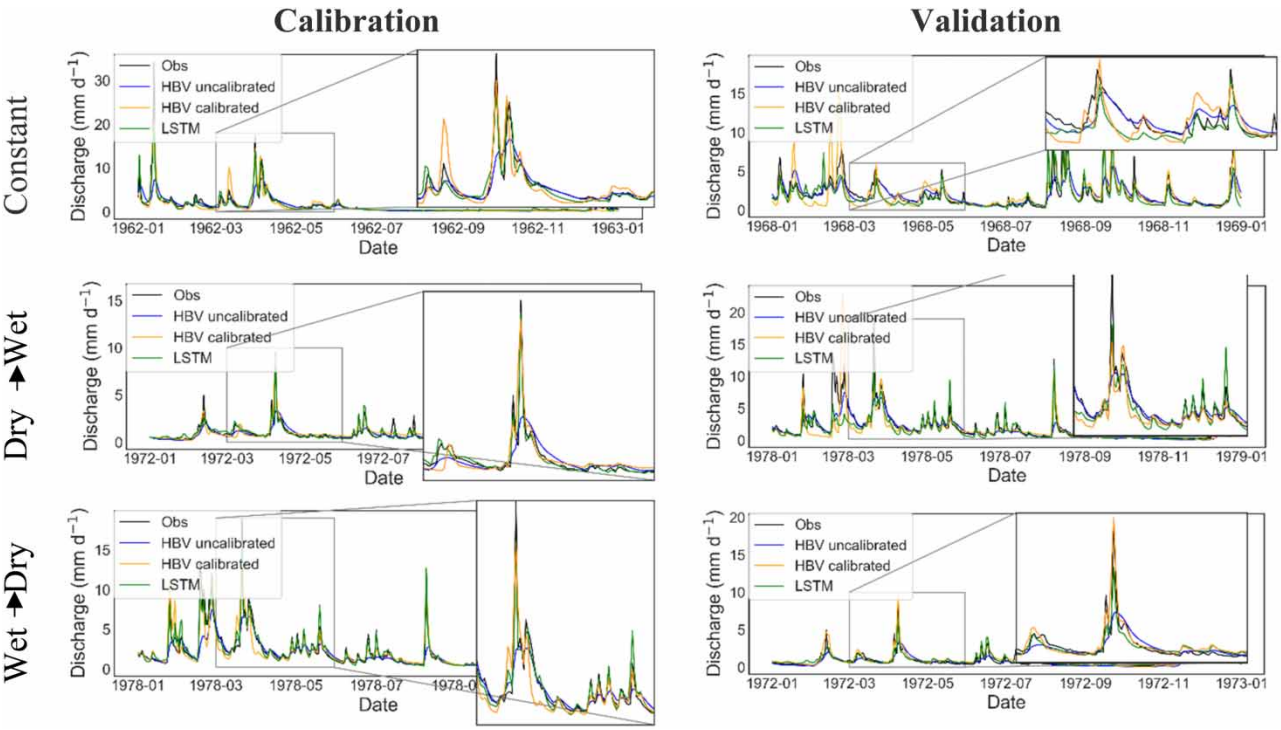

Figure 3 | Observed and simulated hydrographs.

Overall, when calibrated, the LSTM resulted in a much better fit than the HBV. However, the performance drop when going into a validation mode is also much larger for the LSTM (less robust). For the validation period, the LSTM was at best as good as the HBV (especially for other criteria than used in calibration and for changing conditions).

The LSTM was shown to be more dependent on dataset length to perform as well as the bucket-type model. The improvement in model performance/robustness with the increase of the time-series size used in calibration was also observed by Ayzel & Heistermann (2021) and Gauch *et al.* (2019) while testing the performance of the LSTM networks for streamflow prediction in constant conditions. Here, this positive correlation was more pronounced for the constant conditions than that for changing conditions. The lesser contribution of the time-series size in model performance under changing conditions may be explained by the limitation of the data provided for the calibration (only dry or wet periods used in calibration), that is, less information about the hydrological processes was provided to the model. The physical constraints of the HBV model made the need for longer data series in calibration less important, indicating the suitability of this model for predictions when data are limited. The same was observed by Ayzel & Heistermann (2021) while comparing an LSTM network to the GR4H conceptual model.

The robustness analysis showed that the LSTM is robust enough for climate transposability to a drier period. The generalization from a dry period to a wetter period is less satisfactory, mainly because in this case, the model needs to extrapolate to a discharge range not used in calibration, as also reported by Pan *et al.* (2019) and Wilby (2005) employing traditional hydrological models.

It is important to highlight that in this model setup, the LSTM was trained on individual catchments and its calibration over a large number of catchments and with a larger data series can yield better results and should be explored further. However, comparing models with different structures is not an easy task, especially when trying to keep a fair comparison between the models. More sophisticated hyperparameter tuning techniques may also improve the LSTM model's simulations, as well as coupling the model with process-based models.

CONCLUSION

In this work, we tested the predictive ability of the LSTM for daily discharge prediction in snow-influenced catchments under changing conditions. When calibrated, the LSTM resulted in a much better fit than the HBV, however, in a validation mode,

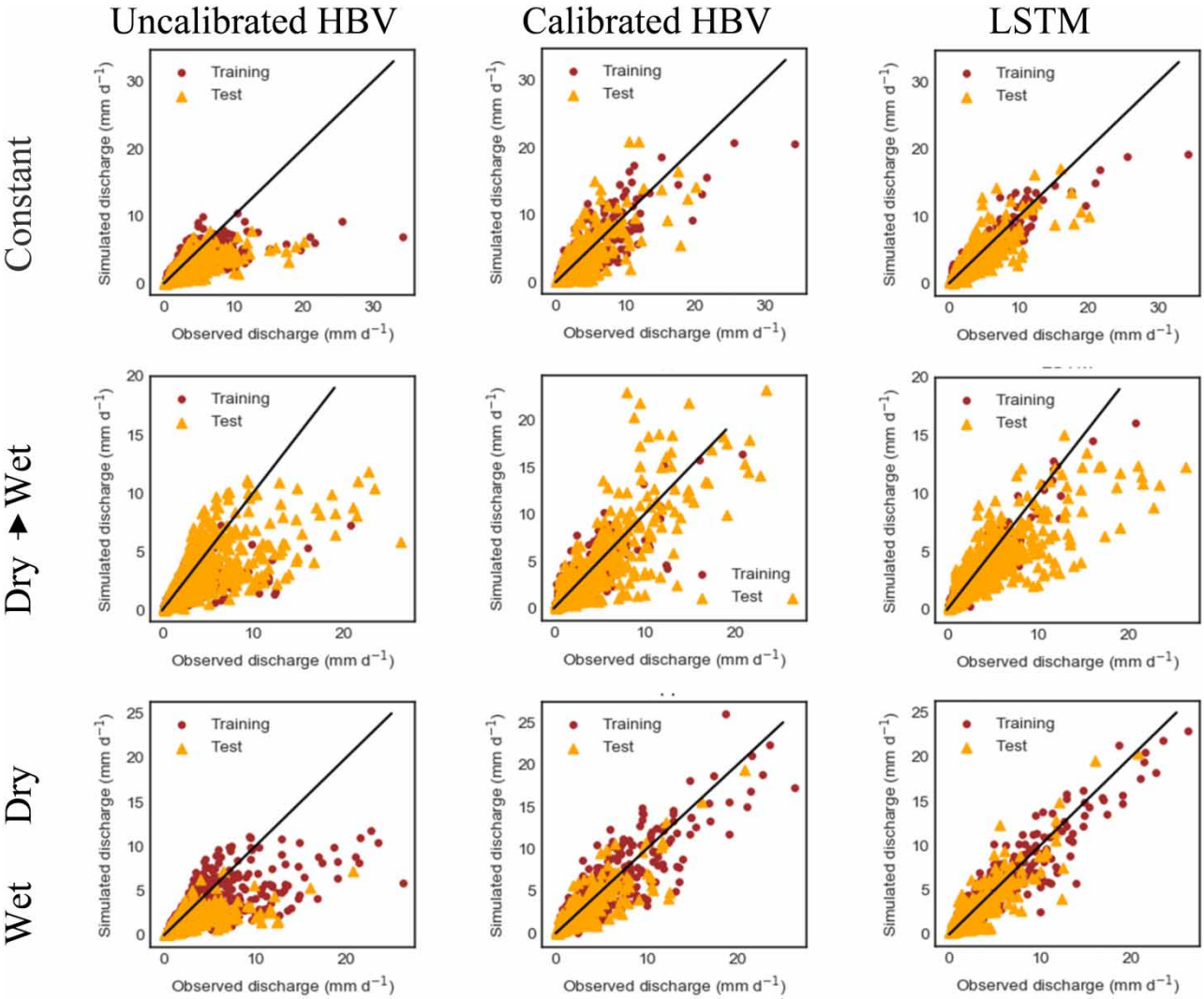

Figure 4 | Scatter plot of the observed and the simulated discharge.

Table 2 | Average robustness of the LSTM (left number) and the calibrated HBV model (right number) defined by the subtraction between the efficiency in calibration and validation

		Constant			
		NSE	**KGE**	**NPE**	**MARE**
Calibration period length (years)	2	0.28|0.04	**0.16|0.03**	0.14|0.01	0.27|0.06
	3	0.16|0.00	0.17|0.00	0.07| − 0.01	0.1| − 0.02
	4	0.12|0.02	**0.09|0.04**	**0.07|0.03**	**0.08|0.04**
	6	**0.14|0.10**	**0.10|0.04**	**0.08|0.07**	**0.12|0.07**
	Dry → Wet				
	2	**0.28|0.13**	**0.17|0.04**	0.14|0.04	**0.18|0.00**
	3	**0.24|0.13**	0.21|0.08	**0.13|0.05**	0.06| − 0.06
	4	0.26|0.09	0.18|0.02	0.11| − 0.01	0.06| − 0.18
	6	**0.19|0.15**	0.14|0.04	**0.05|0.00**	0.07| − 0.04
	Wet → Dry				
	2	**0.31|0.24**	**0.28|0.20**	0.23|0.03	0.54|0.14
	3	**0.31|0.09**	**0.21|0.07**	0.21|0.05	**0.34|0.18**
	4	**0.32|0.18**	**0.22|0.14**	0.17|0.07	0.41|0.20
	6	0.13|0.04	**0.12|0.07**	**0.11|0.06**	**0.16|0.10**

Bold values are showed when the LSTM did not differ significantly from the calibrated HBV. Underlined values are showed when the LSTM under a changing condition did not differ significantly from the constant condition ($\alpha = 0.05$).

the LSTM often performed worse than the HBV (especially for other criteria than used in calibration and for changing conditions). The performance drop when going into a validation mode was larger for the LSTM, indicating less robustness, and the data-driven model was shown to be more dependent on dataset length used in calibration to deliver robustness comparable to a bucket-type model.

Despite this, the results indicate that using longer data series in calibration can benefit the use of the LSTM in contrasting conditions. We recommend that other LSTM modeling setups should be studied further to improve the model performance in such conditions.

ACKNOWLEDGEMENTS

The authors thank the research funding agencies CAPES-Brazil and ESKAS – Swiss Government Excellence Scholarship for the scholarships granted to the first author at different periods of time during this research.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

REFERENCES

Ayzel, G. & Heistermann, M. 2021 The effect of calibration data length on the performance of a conceptual hydrological model versus LSTM and GRU: a case study for six basins from the CAMELS dataset. *Computers & Geosciences* **149**, 104708. https://doi.org/10.1016/j.cageo.2021.104708.

Bastola, S., Murphy, C. & Sweeney, J. 2011 The role of hydrological modelling uncertainties in climate change impact assessments of Irish river catchments. *Advances in Water Resources* **34**(5), 562–576. https://doi.org/10.1016/j.advwatres.2011.01.008.

Blöschl, G., Bierkens, M. F., Chambel, A., Cudennec, C., Destouni, G., Fiori, A. *et al.* 2019 Twenty-three unsolved problems in hydrology (UPH) – a community perspective. *Hydrological Sciences Journal* **64**(10), 1141–1158. https://doi.org/10.1080/02626667.2019.1620507.

Brigode, P., Oudin, L. & Perrin, C. 2013 Hydrological model parameter instability: a source of additional uncertainty in estimating the hydrological impacts of climate change? *Journal of Hydrology* **476**, 410–425. https://doi.org/10.1016/j.jhydrol.2012.11.012.

Broderick, C., Matthews, T., Wilby, R. L., Bastola, S. & Murphy, C. 2016 Transferability of hydrological models and ensemble averaging methods between contrasting climatic periods. *Water Resources Research* **52**(10), 8343–8373. https://doi.org/10.1002/2016WR018850.

Chiew, F. H. S., Teng, J., Vaze, J., Post, D. A., Perraud, J. M., Kirono, D. G. C. & Viney, N. R. 2009 Estimating climate change impact on runoff across southeast Australia: method, results, and implications of the modeling method. *Water Resources Research* **45**(10). https://doi.org/10.1029/2008WR007338.

Coron, L., Andreassian, V., Perrin, C., Lerat, J., Vaze, J., Bourqui, M. & Hendrickx, F. 2012 Crash testing hydrological models in contrasted climate conditions: an experiment on 216 Australian catchments. *Water Resources Research* **48**(5). doi:10.1029/2011wr011721.

Dakhlaoui, H., Ruelland, D., Tramblay, Y. & Bargaoui, Z. 2017 Evaluating the robustness of conceptual rainfall-runoff models under climate variability in northern Tunisia. *Journal of Hydrology* **550**, 201–217. https://doi.org/10.1016/j.jhydrol.2017.04.032.

Dawson, C. W. & Wilby, R. 1998 An artificial neural network approach to rainfall-runoff modelling. *Hydrological Sciences Journal* **43**(1), 47–66. https://doi.org/10.1080/02626669809492102.

Dibike, Y. B. & Solomatine, D. P. 2001 River flow forecasting using artificial neural networks. *Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere* **26**(1), 1–7. https://doi.org/10.1016/S1464-1909(01)85005-X.

Fowler, K. J. A., Peel, M. C., Western, A. W., Zhang, L. & Peterson, T. J. 2016 Simulating runoff under changing climatic conditions: revisiting an apparent deficiency of conceptual rainfall-runoff models. *Water Resources Research* **52**, 1820–1846. https://doi.org/10.1002/2015WR018068.

Gauch, M., Mai, J. & Lin, J. 2021 The proper care and feeding of CAMELS: How limited training data affects streamflow prediction. *Environmental Modelling & Software* **135**, 104926. https://doi.org/10.1016/j.envsoft.2020.104926.

Grusson, Y., Anctil, F., Sauvage, S. & Sánchez Pérez, J. M. 2017 Assessing the climatic and temporal transposability of the SWAT model across a large contrasted watershed. *Journal of Hydrologic Engineering* **22**(6), 04017004. doi:10.1061/(ASCE)HE.1943-5584.0001491.

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. https://doi.org/10.1016/j.jhydrol.2009.08.003.

Hallouin, T., Bruen, M. & O'Loughlin, F. E. 2020 Calibration of hydrological models for ecologically relevant streamflow predictions: a trade-off between fitting well to data and estimating consistent parameter sets? *Hydrology and Earth System Sciences* **24**(3), 1031–1054. https://doi.org/10.5194/hess-24-1031-2020.

Her, Y., Yoo, S. H., Cho, J., Hwang, S., Jeong, J. & Seong, C. 2019 Uncertainty in hydrological analysis of climate change: multi-parameter vs. multi-GCM ensemble predictions. *Scientific Reports* **9**(1), 1–22. https://doi.org/10.1038/s41598-019-41334-7.

Hochreiter, S. & Schmidhuber, J. 1997 LSTM can solve hard long time lag problems. In: *Advances in Neural Information Processing Systems*, pp. 473–479. https://doi.org/10.3390/w10111543.

Hu, C., Wu, Q., Li, H., Jian, S., Li, N. & Lou, Z. 2018 Deep learning with a long short-term memory networks approach for rainfall-runoff simulation. *Water* **10**(11), 1543. https://doi.org/10.3390/w10111543.

Klemeš, V. 1986 Operational testing of hydrological simulation models. *Hydrological Sciences Journal* **31**(1), 13–24. https://doi.org/10.1080/02626668609491024.

Kling, H., Stanzel, P., Fuchs, M. & Nachtnebel, H.-P. 2015 Performance of the COSERO precipitation–runoff model under non-stationary conditions in basins with different climates. *Hydrological Sciences Journal* **60**, 1374–1393. https://doi.org/10.1080/02626667.2014.959956.

Köplin, N., Schädler, B., Viviroli, D. & Weingartner, R. 2014 Seasonality and magnitude of floods in Switzerland under future climate change. *Hydrological Processes* **28**(4), 2567–2578. https://doi.org/10.1002/hyp.9757.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K. & Herrnegger, M. 2018 Rainfall–runoff modelling using long short-term memory (LSTM) networks. *Hydrology and Earth System Sciences* **22**(11), 6005–6022. https://doi.org/10.5194/hess-22-6005-2018.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S. & Nearing, G. 2019a Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences* **23**(12). https://doi.org/10.5194/hess-23-5089-2019.

Kratzert, F., Herrnegger, M., Klotz, D., Hochreiter, S. & Klambauer, G. 2019b Neural hydrology – interpreting LSTMs in hydrology. In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science*, Vol. 11700 (Samek, W., Montavon, G., Vedaldi, A., Hansen, L. & Müller, K. R., eds.). Springer, Cham. https://doi.org/10.1007/978-3-030-28954-6_19.

Lee, G., Jung, S. & Lee, D. 2018 Comparison of physics-based and data-driven models for streamflow simulation of the Mekong river. *Journal of Korea Water Resources Association* **51**(6), 503–514.

Lees, T., Buechel, M., Anderson, B., Slater, L., Reece, S., Coxon, G. & Dadson, S. J. 2021 Benchmarking data-driven rainfall-runoff models in Great Britain: a comparison of LSTM-based models with four lumped conceptual models. *Hydrology and Earth System Science Discussions* [preprint]. https://doi.org/10.5194/hess-2021-127, in review.

Li, C. Z., Zhang, L., Wang, H., Zhang, Y. Q., Yu, F. L. & Yan, D. H. 2012 The transferability of hydrological models under nonstationary climatic conditions. *Hydrology and Earth System Science* **16**, 1239–1254. https://doi.org/10.5194/hess-16-1239-2012.

Li, H., Beldring, S. & Xu, C.-Y. 2015 Stability of model performance and parameter values on two catchments facing changes in climatic conditions. *Hydrological Sciences Journal* **60**, 1317–1330. https://doi.org/10.1080/02626667.2014.978333.

Merz, R., Parajka, J. & Blöschl, G. 2011 Time stability of catchment model parameters: implications for climate impact analyses. *Water Resources Research* **47**, W02531. https://doi.org/10.1029/2010WR009505.

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. https://doi.org/10.1016/0022-1694(70)90255-6, 1970.

O., S., Dutra, E. & Orth, R. 2020 Robustness of process-based versus data-driven modeling in changing climatic conditions. *Journal of Hydrometeorology* **21**(9), 1929–1944. https://doi.org/10.1175/JHM-D-20-0072.1.

Orth, R., Staudinger, M., Seneviratne, S. I., Seibert, J. & Zappa, M. 2015 Does model performance improve with complexity? A case study with three hydrological models. *Journal of Hydrology* **523**, 147–159. https://doi.org/10.1016/j.jhydrol.2015.01.044.

Ouermi, K. S., Paturel, J. E., Adounpke, J., Lawin, A. E., Goula, B. T. A. & Amoussou, E. 2019 Comparison of hydrological models for use in climate change studies: a test on 241 catchments in West and Central Africa. *Comptes Rendus Geoscience* **351**(7), 477–486. https://doi.org/10.1016/j.crte.2019.08.001.

Pan, Z., Liu, P., Gao, S., Xia, J., Chen, J. & Cheng, L. 2019 Improving hydrological projection performance under contrasting climatic conditions using spatial coherence through a hierarchical Bayesian regression framework. *Hydrology and Earth System Sciences* **23**(8), 3405–3421. https://doi.org/10.5194/hess-23-3405-2019.

Pool, S., Vis, M. & Seibert, J. 2018 Evaluating model performance: towards a non-parametric variant of the Kling-Gupta efficiency. *Hydrological Sciences Journal* **63**(13–14), 1941–1953. https://doi.org/10.1080/02626667.2018.1552002.

Rafaeli Neto, S. L., Sá, E. A. S., Debastiani, A. B., Padilha, V. L. & Antunes, T. A. 2019 Efficacy of rainfall-runoff models in loose coupling spacial decision support systems modelbase. *Water Resources Management* **33**(3), 889–904. https://doi.org/10.1007/s11269-018-2086-2.

Refsgaard, J. C. & Knudsen, J. 1996 Operational validation and intercomparison of different types of hydrological models. *Water Resources Research* **32**(7), 2189–2202. https://doi.org/10.1029/96WR00896.

Seibert, J. 2003 Reliability of model predictions outside calibration conditions: paper presented at the Nordic Hydrological Conference (Røros, Norway 4–7 August 2002). *Hydrology Research* **34**(5), 477–492. https://doi.org/10.2166/nh.2003.0019.

Seibert, J. & Vis, M. J. 2012 Teaching hydrological modeling with a user-friendly catchment-runoff-model software package. *Hydrology and Earth System Sciences* **16**(9), 3315–3325. https://doi.org/10.5194/hess-16-3315-2012.

Seibert, J., Vis, M. J., Lewis, E. & Meerveld, H. V. 2018 Upper and lower benchmarks in hydrological modelling. *Hydrological Processes* **32**(8), 1120–1125. https://doi.org/10.1002/hyp.11476.

Seiller, G., Anctil, F. & Perrin, C. 2012 Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions. *Hydrology and Earth System Sciences* **16**(4), 1171–1189. https://doi.org/10.5194/hess-16-1171-2012.

Seiller, G., Hajji, I. & Anctil, F. 2015 Improving the temporal transposability of lumped hydrological models on twenty diversified U.S. watersheds. *Journal of Hydrology: Regional Studies* **3**, 379–399. https://doi.org/10.1016/j.ejrh.2015.02.012.

Shen, C. 2018 A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resources Research* **54**(11), 8558–8593. https://doi.org/10.1029/2018WR022643.

Shortridge, J. E., Guikema, S. D. & Zaitchik, B. F. 2016 Machine learning methods for empirical streamflow simulation: a comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds. *Hydrology and Earth System Sciences* **20**(7). doi: 10.5194/hess-20-2611-2016.

Staudinger, M., Stahl, K., Seibert, J., Clark, M. P. & Tallaksen, L. M. 2011 Comparison of hydrological model structures based on recession and low flow simulations. *Hydrology and Earth System Sciences* **15**(11), 3447–3459. https://doi.org/10.5194/hess-15-3447-2011.

Thirel, G., Andréassian, V. & Perrin, C. 2015 On the need to test hydrological models under changing conditions. *Hydrological Sciences Journal* **60**(7–8), 1165–1173. doi:10.1080/02626667.2015.1050027.

Vaze, J., Post, D. A., Chiew, F. H. S., Perraud, J. M., Viney, N. R. & Teng, J. 2010 Climate non-stationarity–validity of calibrated rainfall–runoff models for use in climate change studies. *Journal of Hydrology* **394**(3–4), 447–457. https://doi.org/10.1016/j.jhydrol.2010.09.018.

Vormoor, K., Heistermann, M., Bronstert, A. & Lawrence, D. 2018 Hydrological model parameter (in)stability – 'Crash testing' the HBV model under contrasting flood seasonality conditions. *Hydrological Sciences Journal* **63**(991), 1007. https://doi.org/10.1080/02626667.2018.1466056.

Wilby, R. L. 2005 Uncertainty in water resource model parameters used for climate change impact assessment. *Hydrological Processes* **19**(16), 3201–3219. https://doi.org/10.1002/hyp.5819.

Xu, C. Y. 1999 Operational testing of a water balance model for predicting climate change impacts. *Agricultural and Forest Meteorology* **98**, 295–304. doi:10.1016/S0168-1923(99)00106-9.

Xu, W., Jiang, Y., Zhang, X., Li, Y., Zhang, R. & Fu, G. 2020 Using long short-term memory networks for river flow prediction. *Hydrology Research* **51**(6), 1358–1376. https://doi.org/10.2166/nh.2020.026.

Zhu, Q., Zhang, X., Ma, C., Gao, C. & Xu, Y. P. 2016 Investigating the uncertainty and transferability of parameters in SWAT model under climate change. *Hydrological Sciences Journal* **61**(5), 914–930. https://doi.org/10.1080/02626667.2014.1000915.

First received 4 May 2021; accepted in revised form 10 March 2022. Available online 8 April 2022