https://doi.org/10.5194/hess-2023-168
Preprint. Discussion started: 26 July 2023

Hydrology and Earth System Sciences
Discussions

Metamorphic Testing of Machine Learning and Conceptual Hydrologic Models

Peter Reichert[1,2], Kai Ma[3,4], Marvin Höge[1], Fabrizio Fenicia[1], Marco Baity-Jesi[1], Dapeng Feng[5], and Chaopeng Shen[5]

[1]Eawag: Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland
[2]current status: retired from Eawag, see https://peterreichert.github.io for updated information
[3]Institute of International Rivers and Eco-Security, Yunnan University, Kunming, China
[4]Yunnan Key Laboratory of International Rivers and Transboundary Eco-security, Yunnan University, Kunming, China
[5]Civil and Environmental Engineering, Pennsylvania State University, University Park, PA, USA

Correspondence: peter.reichert@emeriti.eawag.ch; https://peterreichert.github.io

Abstract. Predicting the response of hydrologic systems to modified driving forces, beyond patterns that have occurred in the past, is of high importance for estimating climate change impacts or the effect of management measures. This kind of predictions requires a model, but the impossibility of testing such predictions against observed data makes it still difficult to estimate their reliability. Metamorphic testing offers a methodology for assessing models beyond validation with real data. It consists of defining input changes for which the expected responses are assumed to be known at least qualitatively, and to test model behavior for consistency with these expectations. To increase the gain of information and reduce the subjectivity of this approach, we extend this methodology to a multi-model approach and include a sensitivity analysis of the predictions to training or calibration options. This allows us to quantitatively analyse differences in predictions between different model structures and calibration options in addition to the qualitative test to the expectations. In our case study, we apply this approach to selected conceptual and machine learning hydrological models calibrated to basins from the CAMELS data set. Our results confirm the superiority of the machine learning models over the conceptual hydrologic models regarding the quality of fit during calibration and validation periods. However, we also find that the response of machine learning models to modified inputs can deviate from the expectations and the magnitude and even the sign of the response can depend on the training data. In addition, even in cases in which all models passed the metamorphic test, there are cases in which the quantitative response is different for different model structures. This demonstrates the importance of this kind of testing beyond the usual calibration-validation analysis to identify potential problems and stimulate the development of improved models.

1 Introduction

The availability of hydrologic and meteorological data and catchment attributes for a large number of catchments in the USA (Newman et al., 2015; Addor et al., 2017) has greatly stimulated hydrologic research in the past few years. In particular, it has been shown that the training of machine learning models jointly to hydrologic data from a large number of catchments leads to an extraordinary performance of these models even for the prediction of the output of catchments that had not been used for training (Kratzert et al., 2018, 2019a, b; Feng et al., 2020, 2021). Arguably, this breakthrough was made possible by the combination of two elements:

(i) using (machine learning) models that are highly flexible and contain a large number of parameters;

(ii) training the models jointly on large sets of diverse basins using relevant catchment attributes as additional input to meteorological time series to allow the models to learn diverse response patterns and their dependence on catchment characteristics.

Due to the use of a large and diverse data set, overfitting of the models is mitigated and the models to some degree gain the capability of acquiring hydrologic knowledge. It has been shown that this kind of hydrologic knowledge can even be transferred across continents (Ma et al., 2021). Shen (2018) provides a trans-disciplinary review of deep learning research and its relevance for water resources research. The availability of many more data sets for other countries, such as Chile (Alvarez-Garreton et al., 2018), Great Britain (Coxon et al., 2020), Brazil (Chagas et al., 2020), Australia (Fowler et al., 2021), and more, bears great potential for further development of hydrologic modelling across catchments. The success of a large number of machine learning-based studies using such data sets has challenged the belief of hydrologists that the prediction of the output of ungauged catchments would only be possible with models that are built with strong support by hydrologic expert knowledge (Hrachowitz et al., 2013; Nearing et al., 2021).

The primary focus of the studies cited above was so far on model training and validation on a future part of the time series or on catchments not used for calibration. The question whether this success is transferable to the prediction of the consequences of modified driving forces in these catchments is less certain (Bai et al., 2021; Natel de Moura et al., 2022). For the prediction of the effects of climate change and water management measures on the hydrology of catchments, it is of particular interest to modify driving forces beyond the patterns observed in the past. When the perturbation is large enough, there is no data available for validating the models under such perturbations. The problem is of different nature for conceptual hydrologic models and for machine learning models. The prediction of the behavior of catchments under modified driving forces with conceptual models is challenging because it is very hard to predict the required modifications to model parameters induced by changes in vegetation, soil, etc. (Merz et al., 2011). The prediction with machine learning models could either lead to wrong results due to poor out-of-domain generalization (Wang et al., 2022), or the results could be much better due to a more comprehensive consideration of adapted catchment properties learned from other catchments in the training set. Which of these effects dominates may depend on the degree of input modifications and on the diversity of the set of basins used for training. For these reasons,

https://doi.org/10.5194/hess-2023-168
Preprint. Discussion started: 26 July 2023

it is of interest to compare predictions of both kinds of models under modified driving forces and to investigate whether the results depend on the training data set and on parameters of the optimization algorithm.

It is the goal of this study to compare the behavior of machine-learning and conceptual models under modified driving forces and to investigate to which degree we can learn about deficiencies of models and pathways for their improvement from these results. We will do this by considering isolated changes in precipitation and temperature and compare the outlet discharge with expected outcomes for selected basins from the CAMELS data set (Newman et al., 2015; Addor et al., 2017) with different characteristics (low-elevation, warm catchments vs. very high elevation, snow-dominated catchments). Note that this is a metamorphic testing design (Xie et al., 2011; Yang and Chui, 2021) that facilitates the formulation of the qualitative expected behavior, rather than a realistic climate change scenario that would consist of coupled temperature and precipitation changes. Based on this design, the more specific goals of our study are to answer the following questions:

1. Are good fits during calibration and validation periods sufficient to gain confidence in predictions under modified driving forces?
2. How useful is metamorphic testing of models beyond the usual calibration-validation analysis?
3. Do machine learning models always improve when extending the training data set?
4. How do machine learning and conceptual models complement each other in terms of strengths and deficits?

2 Methods

2.1 Metamorphic Testing

Metamorphic testing is a methodology for assessing models beyond validation with real data (Xie et al., 2011; Yang and Chui, 2021). It consists of

(i) defining changes to model input for which the expected response of the underlying system is assumed to be known at least qualitatively, and

(ii) testing the model response to these changes for consistency with these expectations.

Note that metamorphic testing does not replace calibration and validation but it is an additional, complementary test to the quality of fit that is specifically targeted at situations (inputs) for which there are no response data available. The input changes underlying metamorphic testing should be designed in such a way that they reflect aspects of inputs that are of interest for predicted output but that they still allow for a qualitative characterization of expected responses. One methodology to design such input changes is to reduce the dimension of the problem by modifying just one input with a relatively simple pattern rather than using correlated input changes in multiple inputs and complicated temporal pattern as it would be needed for real predictions. Such additional tests to model fit are important as it has been shown that quality of fit and prediction accuracy do not necessarily improve in parallel. At least one case study came to the conclusion: "Surprisingly, the prediction accuracy of a model and its ability to provide consistent predictions were found to be uncorrelated" (Yang and Chui, 2021). The conclusions may not always be that extreme, but such cases indicate the need for model testing beyond the quality of fit.

The weakness of metamorphic testing is that it requires the specification of the expected response of a system under modified inputs. Even if we define simple input changes to facilitate the fulfillment of this requirement, it still requires partly subjective expert judgements that may be biased by the limited mechanistic understanding of the system's function by the experts or, more generally, by the incomplete state of current scientific knowledge. To further increase the understanding of model behavior and reduce the subjectivity of testing, we extend the test by using a multi-model approach to the analysis of the sensitivity of the results to model structure and to different training or calibration options. This allows us to identify quantitative deviations of predictions (to modified inputs) between model structures and sensitivities of the predictions to calibration options in addition to the qualitative metamorphic test.

There are four potential outcomes of this extended metamorphic testing approach:

(A) Metamorphic test succeeded, models mutually consistent. The predicted response is robust against the investigated model structures and changes in the calibration process and agrees with the expectations. This result confirms the model structure and increases the trust into reliable predictions.

(B) Metamorphic test succeeded, some models inconsistent with others. The predicted response is weakly sensitive to the investigated model structures and aspects of the model calibration process in the sense that there are significant changes in the response but they are still within the range of the qualitative expert predictions. This result shows the limits of metamorphic testing but the identified differences between responses may still stimulate thinking about model structure improvements.

(C) Metamorphic test failed, some models inconsistent with others. The predicted response is highly sensitive to the investigated model structures or aspects of the model calibration process with responses in agreement and in disagreement with the expert expectations. This indicates problems of the models to reliably predict the response to the investigated input changes and indicates the need for a revision of model structures.

(D) Metamorphic test failed, models mutually consistent. The predicted response is robust against the investigated model structures and changes in the calibration process but it disagrees with the expectations. This clearly demonstrates a serious problem either caused by deficits of the model structure that lead to wrong predictions or to incomplete scientific knowledge that lead to incorrect expert predictions. This is the most difficult outcome of the metamorphic analysis but it still demonstrates its importance as it uncovered a problem. In this case it is very important to think of potential mechanisms that may have been overlooked by the experts as well as structural deficits of the investigated models. This may initiate an extended research process that depends on the investigated system and models.

Having climate change predictions in mind but focusing on more simple input changes for model testing, for metamorphic testing we chose an isolated precipitation change and an isolated temperature change, both with simple temporal pattern. As mentioned before, this setup covers inputs relevant for climate change predictions, but is intended to represent simpler input changes for which it may be easier for experts to characterize the response compared to realistic input changes for climate change. In the following, we describe these input changes and the expected responses:

1. *Input change*: **Constant relative increase in precipitation by 10%.**
   Investigated response: We are interested in the change in discharge at the catchment outlet resulting from the change in precipitation:

   $$\Delta Q_P = Q(1.1 \cdot P,T) - Q(P,T) \quad , \tag{1}$$

   where Q is the hydrologic model describing catchment outlet discharge as a function of precipitation time series, P, and temperature time series, T. ΔQ_P is the change in catchment outlet discharge resulting from the 10% increase in precipitation as predicted by the model.
   Expected response: We expect an increase in catchment outlet discharge that reflects the discharge pattern of the base simulation. Only in cases of short events and considerable travelling of the flood wave, we expect a decrease in discharge at the falling limb of the discharge peak (following an increase at the rising limb), due to a shift of the flood peak to earlier times caused by a higher flood wave celerity at higher water levels (Battjes and Labeur, 2017). This expectation is based on the assumption that a 10% increase in precipitation is small enough to not fundamentally change vegetation, soil, and other catchment properties. Such changes may lead to a nonlinear response, but we do not expect that they will be strong enough to invert the sign of the response. Note, however, that more complex response patterns are possible for more complex input changes, as discussed by Blöschl et al. (2019). However, also in this study, increases and decreases in flood levels are caused by regionally different changes in precipitation with decreased precipitation leading to decreased flood levels and increased precipitation to increased flood levels. In a world-wide analysis of past trends in water balance and evapotranspiration, Ukkola and Prentice (2013) found some regions (Europe and Canada) with increasing precipitation and decreasing runoff (see Figure 5 in that paper). However, as this is an analysis of past data, many other factors changed also, in particular, there was a significant temperature increase in these regions that contributed to increased evapotranspiration whereas we assume no change in temperature for this input change scenario.

2. *Input change:* **Constant increase in temperature by 1°C.**
   Investigated response: We are interested in the change in discharge at the catchment outlet resulting from the change in temperature:

   $$\Delta Q_T = Q(P,T + 1^{\circ}\mathrm{C}) - Q(P,T) \quad , \tag{2}$$

   where ΔQ_T is the change in catchment outlet discharge resulting from the 1°C increase in temperature as predicted by the model and the other symbols have the same meaning as in equation (1).

Expected response: For warm catchments (without snow cover), we expect a decrease in outlet discharge that is more pronounced in summer than in winter due to increased evapotranspiration. Again, in cases of short events and considerable travelling of the flood wave, we may get a short increase in discharge at the falling limb of the peak (following a decrease at the rising limb) due to a shift of the peak to later times caused by a lower flood wave celerity at lower water levels (Battjes and Labeur, 2017). For catchments with seasonal snow cover dynamics dominated discharge, we expect an increase in river discharge in autumn or winter due to a later change of precipitation from rain to snowfall and an earlier melting in spring followed by a decrease in river discharge because the snow melt will terminate earlier. These changes should be more pronounced during rain and snow melt dominated periods than during periods in which a large fraction of precipitation falls as snow even under the warmer temperature regime and turn to the pattern described above for warm catchments after the snow melt period ended and snowfall did not yet start in autumn. There is less empirical evidence for this expected response in past data (Ukkola and Prentice, 2013) because in most regions temperature increase is accompanied by precipitation increase and thus leads to increased discharge. However, there are some cases, particularly in North-Asia (see Figure 5 in Ukkola and Prentice (2013)), where there is increase in temperature and runoff despite no significant trend in precipitation. This may be a consequence of a change in snow cover and vegetation.

If the training data contained precipitation or temperature related catchment attributes, we compared the results to training with omission of these attributes to avoid biased results due to inconsistent changes in driving forces.

The intent of our study is to identify potential problems of hydrologic models to learn from them and not to provide a representative overview of results of different models. For this reason, we select catchments that allow us to test the response pattern described above as well as possible. As finding reasons for poor fit is a complementary technique of improving models on which we do not focus in this paper, we only select catchments for which all our primary modelling approaches lead to a very good fit (NSE > 0.8 during the calibration period for all investigated model structures). All of these models also lead to a good fit during the validation period. To best represent the conditions for which we can describe the expected response as described above, we choose:

- *Low altitude warm basins:*
  These basins should only have a minor amount of snow and thus a relatively simple response patterns as described above.
- *Very high altitude (cold) basins:*
  In these basins, the response should be dominated by the shifts in snowfall and snow melt.

To complement our study, we also chose intermediate altitude catchments:

- *Intermediate altitude basins:*
  For these basins, we expect a combination of the snow-cover-dominated response in winter and spring combined with the warm-basin response in summer. The transition between the two regimes will depend on altitude and latitude, which makes the response less clear than in the other two cases.

Hydrology and Earth System Sciences
Open Access
Discussions

2.2 Models

2.2.1 Conceptual Hydrologic Models

We will compare the conceptual hydrologic model GR4 (Santos et al., 2018), which is a continuous-time version of the model GR4J (Perrin et al., 2003) in combination with a continuous-time version of the snow accumulation model Cemaneige (Valery et al., 2014), which we call "GR4neige", and a continuous-time version of the discrete-time model HBV (Bergström, 1992; Lindström et al., 1997; Seibert, 1999; Seibert and Vis, 2012). All equations of these conceptual hydrologic models are given in Appendix A.

2.2.2 Machine Learning Models

The great success of machine learning in hydrology is primarily based on the Long Short-Term Memory (LSTM) models (Kratzert et al., 2018, 2019a, b; Feng et al., 2020). We will thus also exclusively use the LSTM approach to represent machine learning models. The models deviate from each other by their consideration of basins for calibration and by the set of catchment attributes used for calibration (see section 2.3 below). Appendix B provides an overview of the setup of these models.

2.3 Training

The LSTM model was jointly trained to all 671 basins of the US CAMELS data set (https://ral.ucar.edu/solutions/products/camels) using daymet forcing and the catchment attributes listed in Table B1 in the Appendix (Newman et al., 2015; Addor et al., 2017) and maximizing the Nash-Sutcliffe Efficiency (NSE). As we encountered some unexpected responses in the low-altitude basins to a change in temperature (see section 3.2.1 below), additional trainings were done, as described in section 3.3.

The conceptual hydrologic models used daymet altitude band inputs as provided in the CAMELS data, aggregated to a maximum of 5 bands, for catchment-by-catchment calibration by maximization of the posterior with a simple, uncorrelated, normal error model and wide priors. As there is only incomplete banded input data available for the basins 12167000, 12186000 and 12189500, we calibrated the model only for 668 of the 671 basins of the US CAMELS data set.

In both cases, we used the same 15 years for calibration and the same 15 years for validation as in the original publication by Newman et al. (2015) (1980/10/01-1995/09/30 for calibration and 1995/10/01-2010/09/30 for validation).

2.4 Implementation

The conceptual hydrologic models were implemented in Julia (Bezanson et al., 2012, 2017) using the packages DifferentialEquations.jl (Rackauckas and Nie, 2017), ForwardDiff (Revels et al., 2016), Zygote (Innes, 2018) and Optim (Mogensen and

https://doi.org/10.5194/hess-2023-168
Preprint. Discussion started: 26 July 2023

Riseth, 2018).

The LSTM was implemented in Python (Van Rossum and Drake, 2009) using Pytorch (Paszke et al., 2019).

All our code will be made publicly available *(link will be provided)*.

3 Results and Discussion

3.1 Quality of Fit

Fig. 1 provides an overview of the Nash-Sutcliffe Efficiency (NSE) values achieved for the calibration and validation periods for all modelling approaches, for the 668 basins for which also the conceptual models could be calibrated as well as for the 12 basins selected for metamorphic testing (see next section). These results clearly confirm the strength of the LSTM model compared to the conceptual hydrologic models regarding the quality of fit for calibration as well as validation periods. The LSTM has the additional advantage that it generalizes very well to catchments not used for calibration but this feature is not investigated in this paper.

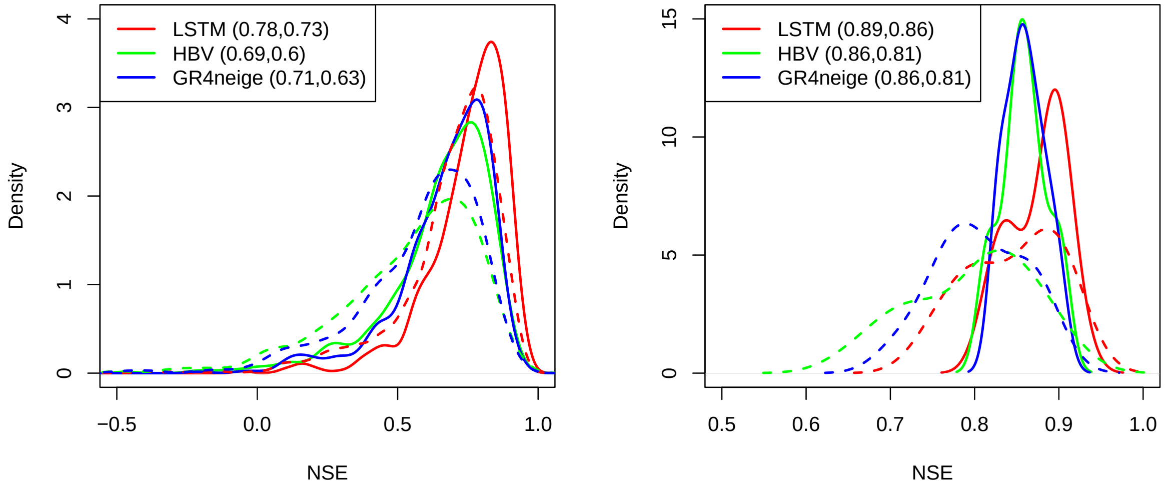

Figure 1. Overview of NSE values of all modelling approaches for the calibration period (solid) and the validation period (dashed) for all 668 basins (left) and for the 12 basins selected for metamorphic testing (right). The median NSE values are indicated in brackets in the legend (calibration period, validation period).

https://doi.org/10.5194/hess-2023-168
Preprint. Discussion started: 26 July 2023

3.2 Metamorphic Testing

For metamorphic testing, we separately evaluated basins that belong to the three classes of low altitude, warm basins, very high altitude basins, and intermediate altitude basins mentioned in section 2. For each of the three classes, we selected four basins. Fig. 2 provides an overview of the locations of the selected four basins within each category. As mentioned in section 2.1 these basins were selected by allowing for an excellent fit for all modelling approaches (NSE > 0.8 during the calibration period). Due to the limited number of basins in these categories, the strong requirement regarding the quality of fit for all modelling approaches, and the wish to have the same number of basins in each category, it was not possible to compare more basins under these constraints. However, as shown in the following sections, there are quite consistent patterns of responses to changes within each of these categories.

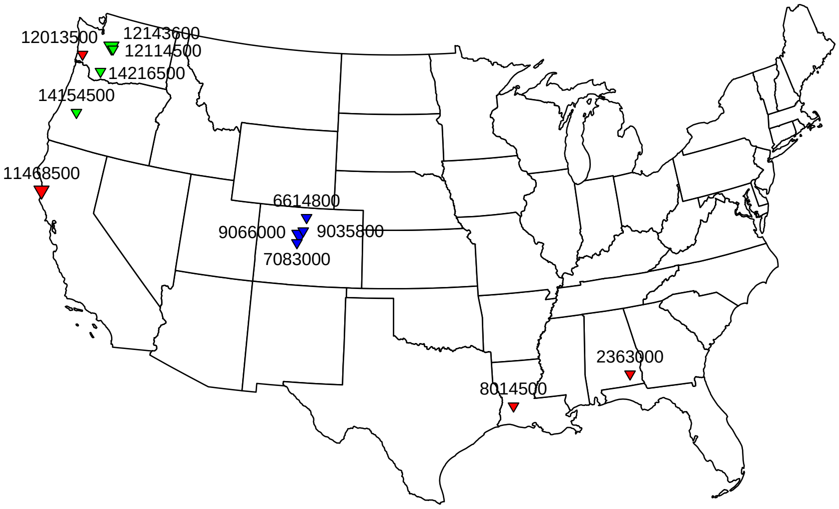

Figure 2. Basins used for metamorphic testing. Red markers indicate low altitude basins, blue markers very high altitude basins, and green markers intermediate altitude basins. The large markers represent basins with results shown in the main paper, the the results for all basins are shown in the Supporting Information.

3.2.1 Low Altitude Warm Basins

Figure 3 shows the results for the final year of the calibration period for a typical warm, low altitude basin. Results for more years during the calibration and validation periods and for more low altitude basins are provided in the Figures SI.1 to SI.16 in the Supporting Information. These results are systematic across all studied basins, demonstrating that the features discussed in this section represent the typical behavior of this kind of basins and are not just an artifact of this specific basin and year.

As is shown by the NSE values in the legends of the fourth panels (Fig. 3 and Figs. SI.1 to SI.16), for these basins, all of the compared primary modelling approaches (GR4neige, HBV, LSTM) provide an excellent fit over the calibration and validation periods (all NSE values are larger than 0.8 during calibration and larger, mostly much larger, than 0.65 during validation, see also the overview of NSE values in Fig. 1).

The sensitivities to a 10 % increase in rainfall, ΔQ_P (see equation 1), are plotted in the top panel of Figure 3 (and of Figs. SI.1 to SI.16). All our modelling approaches (GR4neige, HBV, LSTM) lead to very similar sensitivities to the investigated relative change in precipitation. The sensitivities to the investigated increase in rainfall also correspond to our expectations as described in section 2.1, as they are positive and larger during precipitation events than during dry weather periods (compare time series of the precipitation sensitivities in the top panel to the time series of discharge in the bottom panel). The result of this metamorphic test therefore belongs to the category (A) outlined in section 2.1 (consistent agreement with expectations across modelling approaches) and makes us confident into the response of all models to changes in precipitation.

In contrast to the precipitation sensitivities shown in the first panel, the second panel of Figure 3 (and of Figs. SI.1 to SI.16) shows substantial differences in temperature sensitivities, ΔQ_T (see equation 2), between different modelling approaches (GR4neige, HBV, LSTM). The sensitivities of the hydrologic models GR4neige and HBV are essentially negative with only some brief positive excursions associated with small shifts in discharge peaks. These are the expected sensitivities as described in section 2.1. In contrast, the LSTM often shows a positive response of catchment outlet discharge to the investigated temperature increase, in particular during flood events. This seems to be an implausible response, as increased temperature increases evaporation whereas precipitation does not change in our metamorphic testing scenario. The result of this metamorphic test thus belongs to category (C), outlined in section 2.1 (inconsistency with expectations for some model structures). This raises the question whether the expected response neglects the same processes as the conceptual hydrologic models and the LSTM, through its broad coverage of climatic conditions of 671 Camels basins, better considers the effect of changing catchment properties due to the increasing temperature or if this difference reflects a deficit of the capabilities of the LSTM to predict under changes in driving temperature. Since we here see a striking difference in the behaviour of models that fit and predict very well under current climatic conditions, we have to investigate how consistent the response of the LSTM is across different training options. This can provide additional hints regarding which of the two explanations discussed above may be more plausible. This will be investigated in section 3.3.

3.2.2 Very High Altitude (Cold) Basins

Figure 4 shows the results for the final year of the calibration period for a typical very high altitude, cold basin. Results for more years during the calibration and validation periods and for more high altitude basins are provided in the Figures SI.17 to SI.32 in the Supporting Information. These results demonstrate that the features discussed in this section represent the typical

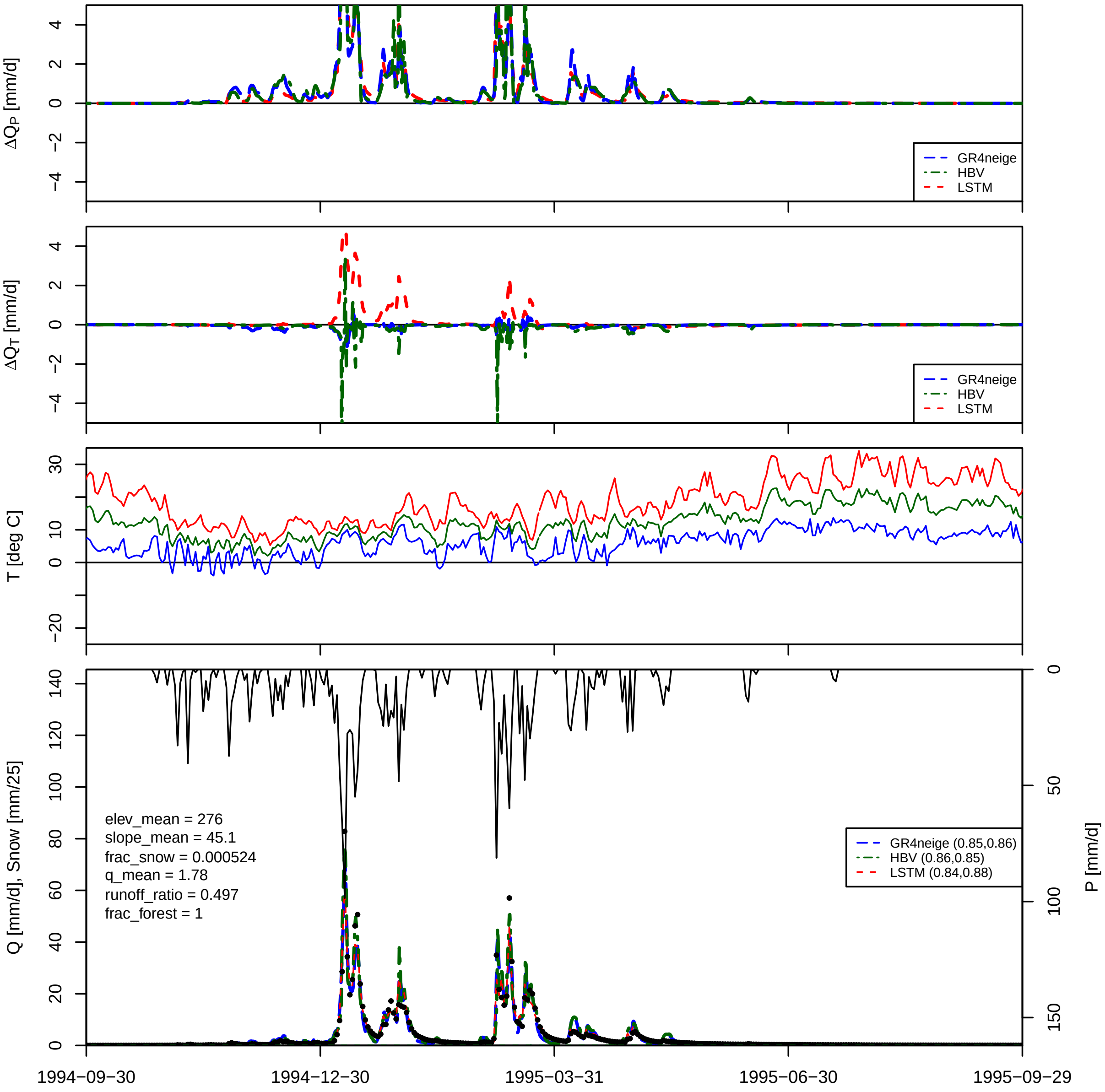

Figure 3. Results for basin 11468500. Top panel: modelled sensitivity of discharge to a 10% increase in precipitation, ΔQ_P (see equation 1). Second panel: modelled sensitivity of discharge to a 1 degree increase in temperature, ΔQ_T (see equation 2). Third panel: minimum (blue), mean (orange) and maximum (red) temperature. Fourth panel: observed precipitation (from top, right axis); modelled and observed discharge and modelled snow cover in max. five altitude bands (bottom, left axis, zero for this particular basin), NSE for calibration and validation periods in brackets in the legend; and the values of selected catchment attributes according to Addor et al. (2017) (on the left).

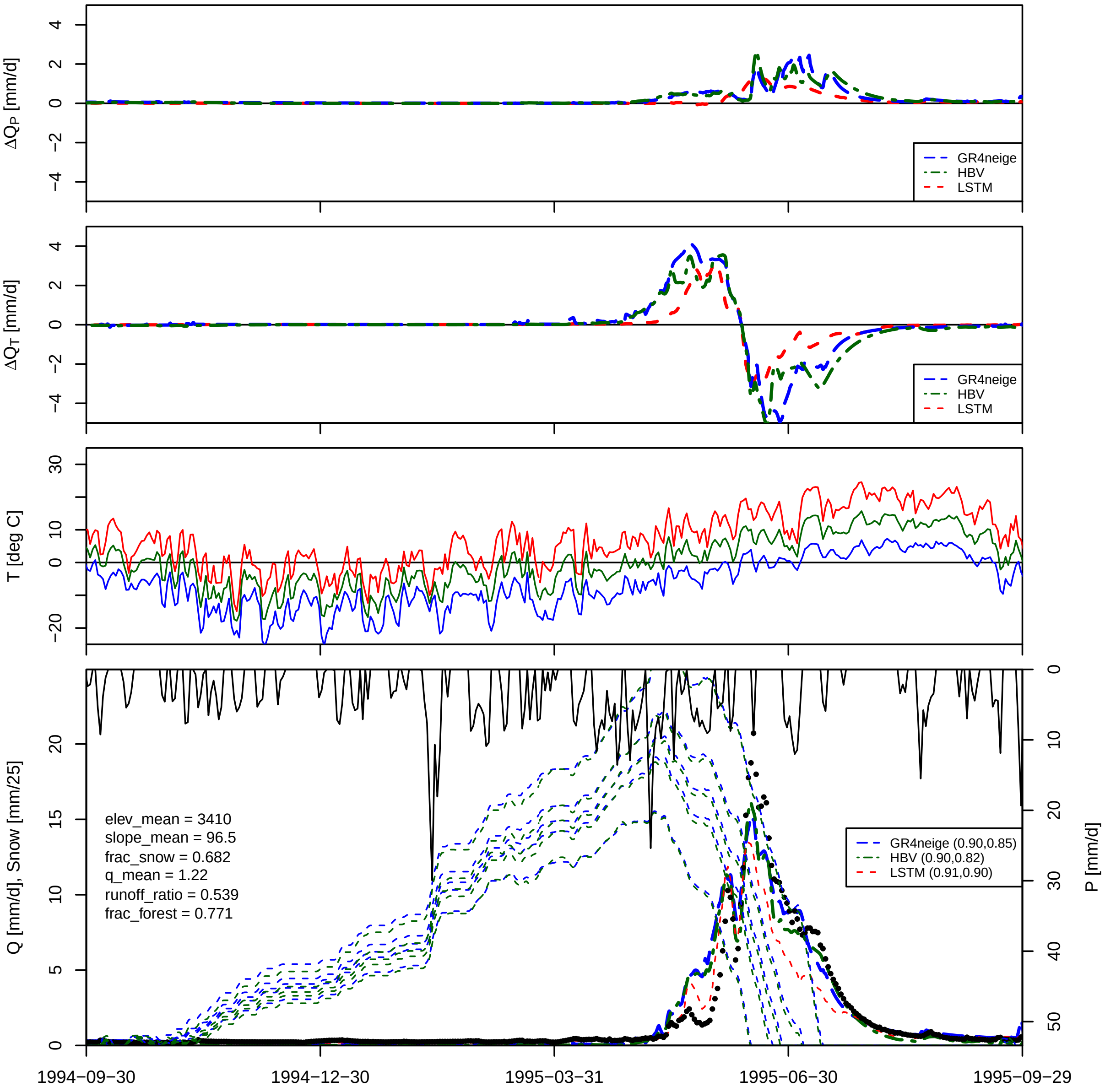

Figure 4. Results for basin 09066000. See Fig. 3 for an explanation of the plot panels.

270 behaviour of this kind of basins and are not just an artefact of this specific basin and year.

The legends of panels 4 in these figures (Fig. 4 and Figs. SI.17 to SI.32) show again that we have an excellent fit during the calibration as well as validation periods for all modelling approaches (GR4neige, HBV, LSTM) with NSE values larger than

0.8 during calibration and larger than 0.7 during the validation periods.

The rainfall sensitivities show in this case more differences as for the low altitude catchments in section 3.2.1. All models show the expected positive rainfall sensitivities, but the response of the LSTM is considerably smaller and smoother than the responses of the conceptual models.

Also the temperature sensitivities show the expected behavior of a positive sensitivity due to the earlier snow melt process followed by a negative sensitivity due to the earlier termination of the snow melt process. There is a tendency that the positive response starts later and the negative response ends earlier for the LSTM compared to the conceptual hydrologic models. Also these sensitivities tend to be smaller and smoother for the LSTM than for the conceptual hydrologic models.

The results of these metamorphic tests thus belong to the category (B) outlined in section 2.1 (significant differences between approaches, but still qualitative agreement with the expectations). As mentioned in section 2.1 this result shows the limits of metamorphic testing, as it is difficult to judge, which of the quantitative responses is closer to reality. Nevertheless, metamorphic testing with multiple models demonstrates that models that provide a similarly good fit during calibration and validation periods can still differ considerably in their response to modified driving forces. This indicates to be cautious with predictions of such responses.

3.2.3 Intermediate Altitude Basins

Figure 5 shows the results for the final year of the calibration period for a typical intermediate altitude basin. Results for more years during the calibration and validation periods and for more intermediate altitude basins are provided in the Figures SI.33 to SI.48 in the Supporting Information. These results demonstrate that the features discussed in this section represent the typical behavior of this kind of basins and are not just an artefact of this specific basin and year.

The legends of panels 4 in these figures (Fig. 5 and Figs. SI.33 to SI.48) show again that we have an excellent fit during the calibration as well as validation periods for all modelling approaches (GR4neige, HBV, LSTM) with NSE values larger than 0.8 during calibration and larger than 0.77 during the validation periods.

The results shown in Fig. 5 combine the results discussed in the previous sections, but resemble more closely the high altitude catchments, as still snow cover dominates the dynamic behavior during most of the season.

The panels 1 and 2 of Fig. 5 show clearly that over the first half of the considered period, all models show very similar responses with respect to the change in precipitation as well as to the change in temperature. In contrast, in the second half of the year, the conceptual models agree with one another but deviate from the LSTM. In this part of the season, the responses of the LSTM is smoother and smaller than that of the hydrologic models. Again, the qualitative nature of metamorphic testing

https://doi.org/10.5194/hess-2023-168
Preprint. Discussion started: 26 July 2023

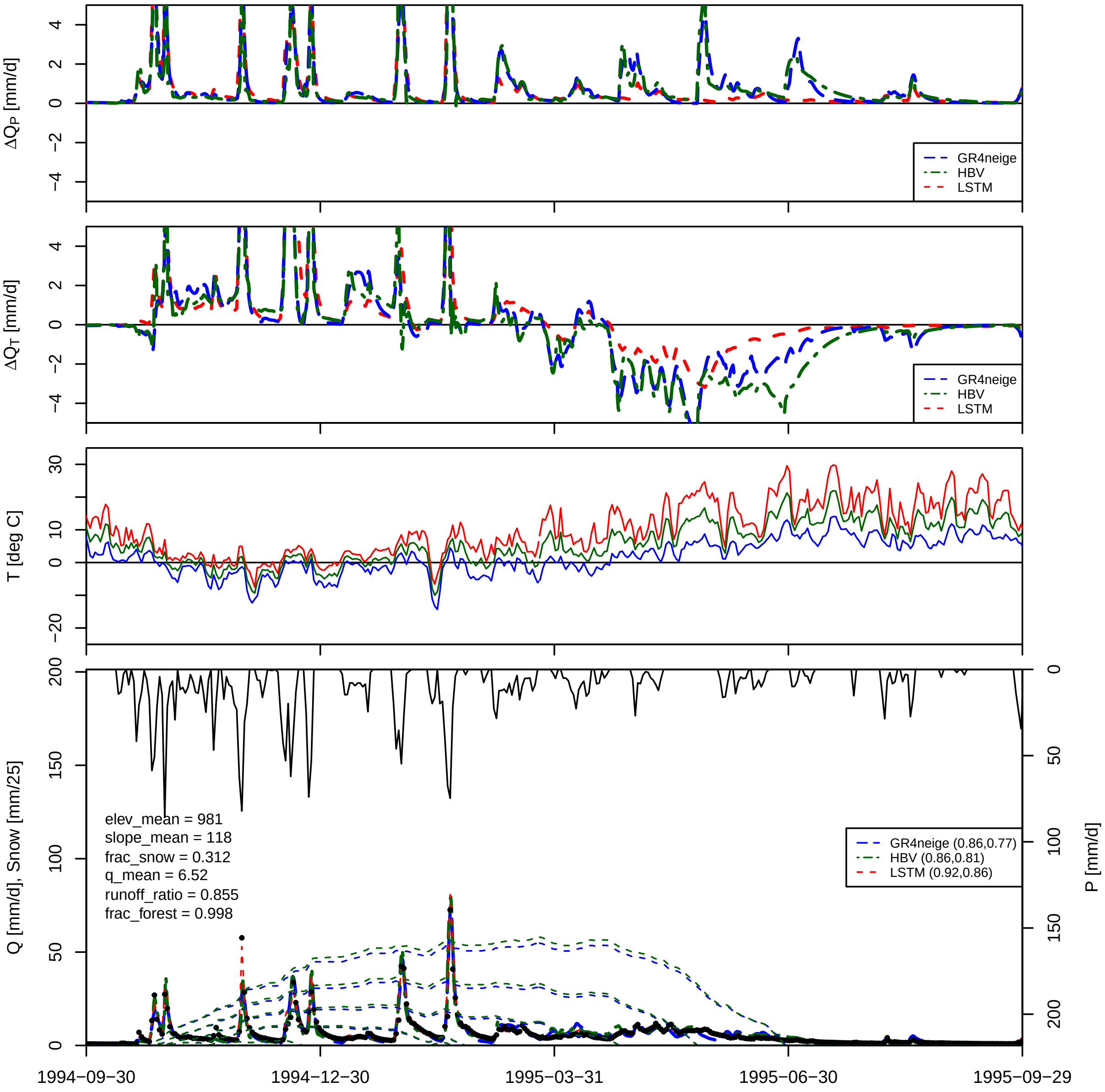

Figure 5. Results for basin 12143600. See Fig. 3 for an explanation of the plot panels.

makes it difficult to assess which of these results are more plausible. These results are again in category (B) of our results classification for metamorphic testing outlined in section 2.1.

https://doi.org/10.5194/hess-2023-168
Preprint. Discussion started: 26 July 2023

3.3 Sensitivity to Attributes, Calibration Set, and Seed for Low Altitude, Warm Basins

As the results for the temperature sensitivities for the low altitude, warm basins were most striking, showing most of the time different signs for the LSTM than for the conceptual hydrologic models, we tried to learn more about the reasons for this phenomenon. To investigate this problem, we performed a sensitivity analysis of the LSTM model regarding

- catchment attributes considered for training,
- basins considered for training, and
- seed of the random number generator that affects the local minimum found by the optimizer.

The idea of using fewer catchment attributes was an attempt to improve the representation of physical processes by the LSTM by only allowing the use of the attributes with a direct physical influence (e.g. omitting mean elevation as temperature has a dominant influence on the physical processes whereas elevation, or air density, is much less relevant) and removing attributes that would have to be modified for prediction with modified input (e.g. all precipitation-related attributes as this information should be inferred from the precipitation time series and would have to be adjusted when modifying precipitation input). The motivation for reducing the set of training basins was to reduce the diversity of basins and primarily keep basins with low elevation (and still sufficient diversity within this class). Finally, the test with different seeds was motivated by checking whether the results where caused by converging into a "bad" minimum, whereas other local minima would lead to better results. Table 1 lists the model versions used for this sensitivity analysis.

Model	Description
GR4neige	conceptual model GR4neige as described in appendix A2
HBV	conceptual model HBV as described in appendix A3
LSTM	LSTM as described in appendix B trained with all 671 basins
LSTM_red	LSTM trained with all 671 basins using only catchment attributes marked with "x" in Table B1
LSTM_515	LSTM trained with the 515 basins with a mean altitude < 1000 m (five different seeds)
LSTM_361	LSTM trained with the 361 basins with a mean altitude < 500 m (five different seeds)
LSTM_211	LSTM trained with the 211 basins with a mean altitude < 300 m (five different seeds)

Table 1. Overview of models. The first three rows describe the basic models planned to be used in the project, the lower four rows are the additional model versions used for the sensitivity analysis to analyze the problem of the deviating temperature sensitivities of the LSTM for low altitude catchments.

Figure 6 shows the precipitation and temperature sensitivities of these models at a higher scale and for a shorter time period than in Figure 3 to facilitate the distinction of the larger number of curves. Results for more years during the calibration and validation periods and for more low altitude basins are provided in the Figures SI.49 to SI.64 in the Supporting Information. The results are qualitatively similar throughout all catchments and periods.

https://doi.org/10.5194/hess-2023-168
Preprint. Discussion started: 26 July 2023

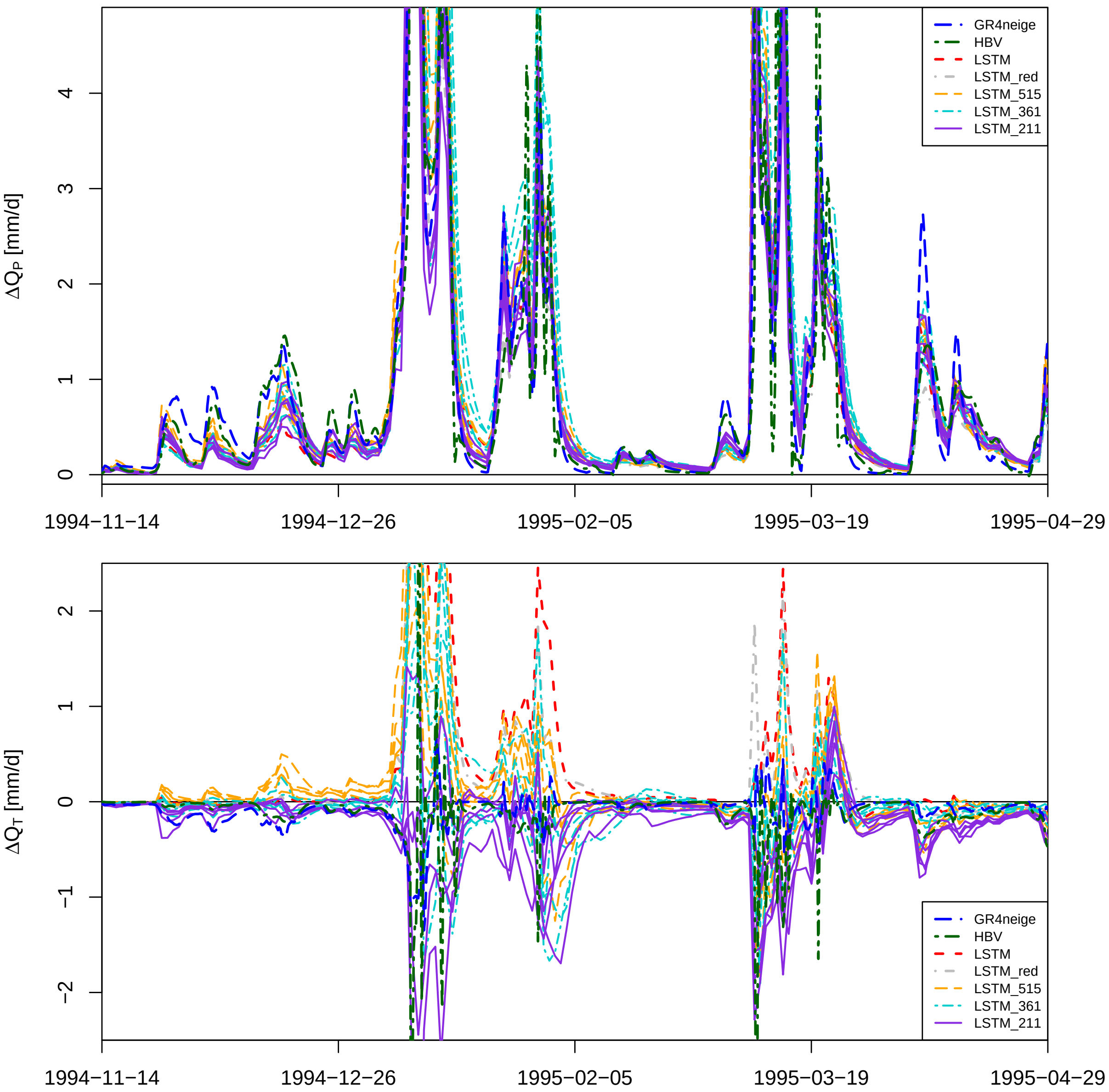

Figure 6. Sensitivities for basin 11468500 and different LSTM calibration options (see Table 1). The sensitivities for the models GR4neige, HBV and LSTM are the same as those shown in Figure 3. In addition, LSTM_red shows the sensitivities when calibrating with a reduced set of catchment attributes, and LSTM_515, LSTM_361 and LSTM_211 show the sensitivities when calibrating the LSTM model with different sets of low-elevation basins. For these three cases, results for 5 different seed values are shown. See text for more details.

The precipitation sensitivities, ΔQ_P (see equation 1), are quite insensitive to any of these modifications from the original setup (see upper panel in Fig. 6 and in Figs. SI.49 to SI.64). In particular, it is remarkable that omitting the mean precipitation

https://doi.org/10.5194/hess-2023-168
Preprint. Discussion started: 26 July 2023

(that had not been changed when increasing the input precipitation time series) from the input does not change the results. This indicates that the response pattern for precipitation change is determined from the input time series rather than from this specific catchment attribute. Also reducing the training data set and changing the random seed does not change the observed precipitation sensitivities.

The results for the temperature sensitivities, ΔQ_T (see equation 2), are quite different (see lower panel in Fig. 6 and in Figs. SI.49 to SI.64). Whereas the reduced set of catchment attributes changes the magnitude, but not the sign of the temperature sensitivities, step-wise reducing the training set to include only lower-level catchments first partly reduces the positive sensitivities (LSTM_515), makes them partly negative (LSTM_361), and, finally, turns them nearly completely negative (LSTM_211). The sensitivities for the LSTM trained only with the basins with mean elevation smaller than 300 m (LSTM_211) are qualitatively similar, but smoother than those from the conceptual hydrologic models. The seed has a quantitative effect, but does not change this qualitative result.

Table 2 summarizes the results of the extended metamorphic testing. Our sensitivity analysis demonstrates that for the LSTM there is a large sensitivity to the training set of modelled catchment outlet response resulting from input temperature changes. It is particularly remarkable, that the sensitivities to temperature change are different despite all of the different calibration versions provide an excellent fit during calibration and validation periods (not shown). This suggests that the problem of making predictions for new environmental conditions, which is very relevant *e.g.* for climate predictions, cannot be detected from only comparing the quality of fit during calibration, validation, and for catchments not used for calibration, as it is has been done in most studies so far. As the responses of LSTM_211 are mostly in qualitative agreement with the expected responses, it seems plausible that the deviations shown for the other calibration options from the expected response are not caused by an error in the expert opinions, but are rather a question of LSTM calibration. This result seems to point that the LSTM trained with all basins is not as universal as expected from previous results, but that an LSTM trained more specifically to the low altitude basins passes the metamorphic test.

4 Summary and Conclusions

We compared the sensitivity of runoff predictions of conceptual and machine learning hydrologic models to changes in precipitation and temperature input for selected catchments of the CAMELS-US data set for which all modelling approaches provided a good fit for the calibration and validation time periods. We found the following results:

- We confirmed earlier results by various researchers that machine learning models provide generally a better fit (higher NSE) for both calibration and validation periods. In addition, machine learning models are much better in extrapolating to basins not used for calibration, but this was not investigated in this study.
- In a metamorphic testing setup, we found qualitatively similar responses of the catchment outlet discharge to precipitation and temperature changes for intermediate- and high-elevation basins, with the main quantitative difference that

Model	low altitude		high altitude		intermed. alt.	
	prec.	temp.	prec.	temp.	prec.	temp.
GR4neige	A	A	B	B	B	B
HBV	A	A	B	B	B	B
LSTM	A	C	B	B	B	B
LSTM_red	A	C				
LSTM_515	A	C				
LSTM_361	A	C				
LSTM_211	A	A				

Table 2. Summary of results of metamorphic testing. See section 2.1 for a description of the result categories (note that the lower four models were designed for low-altitude catchments and were therefore not used for intermediate and high altitude catchments).

the responses of the LSTM were generally smaller and smoother than those of the conceptual hydrologic models. As metamorphic testing is a qualitative procedure, it is hard to assess which of these responses are more plausible. On the other hand, we found major differences in the responses for low altitude basins for which the LSTM led to less plausible results (positive rather than negative responses of catchment outlet discharge to a temperature increase). Training the LSTM with a reduced set of catchment attributes, that should represent the direct physical influence factors, did not resolve this issue. However, training the LSTM on only low-elevation catchments reversed the sign of the sensitivities that then mostly agreed with that of the conceptual hydrologic models. As for the intermediate- and high-elevation basins, the response of the LSTM was then in qualitative agreement with that of the conceptual models but generally smaller and smoother.

These results lead thus to the following conclusions:

A good fit during calibration and validation periods does not guarantee a good response to changes in driving forces. There is a strong need for analyzing model predictions beyond the quality of fit (usually quantified by NSE) and to compare predictions for different model structures to gain confidence in predictions and to gain insight into model prediction uncertainty. This need is evident as we demonstrated that models that fit similarly well for calibration and validation periods can still show strong differences in their response to modified inputs. Good fits during calibration and validation periods are thus not a sufficient criterion for a model to predict the response to modified driving forces accurately. This is a very important conclusion to keep in mind when using hydrologic models of climate change prediction.

Usefulness of metamorphic testing. Metamorphic testing is a very useful tool to test models beyond the usual calibration-validation process. The problem of metamorphic testing is that it requires the response to be at least qualitatively known. This can be difficult and even biased as this requires input of expert knowledge that can be biased by the current state of incomplete scientific knowledge. For this reason, we strongly recommend to go for "extended metamorphic testing" in which we do not only check model predictions for modified inputs with the expected response but in a multi-model approach we also investigate the sensitivity of these results to the model structure and to training data, and algorithmic parameters. This extended analysis can uncover "objective problems" such as a dependence of the response (in our case study even the sign of the response) on choices of the training data, which clearly indicates a problem that is not dependent on the partly subjective prescription of the expected response. It can also, and did in our case study, uncover quantitative differences in responses between different model structures even in cases in which all models passed the qualitative metamorphic test (see the large numbers of "B" classifications in the result Table 2).

Using more data for training can be deleterious. Our results seem to be in contradiction to the general principle that machine learning models always profit from the extension of the training data set. In our case study, adding high elevation basins to the training data set does not reduce the quality of fit but it deteriorates the response of low altitude basins to temperature change. This demonstrates that adding data that is not directly relevant for a specific prediction (in our case for low elevation basins) can have an adverse effect. On the other hand, when we further reduced the training data set, the quality of fit and prediction deteriorated (not shown in the paper). For this reason, this is not a contradiction to the statement that adding "useful data" - data that provides information directly relevant for the question to be investigated - improves the quality of fit and the response to input changes. However, it may raise the awareness for carefully selecting training data as adding less relevant data (for this specific question) may have adverse effects.

Machine Learning vs Conceptual Models. The modelling approaches based on machine learning and on conceptual hydrologic models have complementary strengths and deficits. Machine learning models are particularly strong in providing an excellent quality of fit and prediction accuracy for validation periods as well as for the prediction of ungauged catchments. However, we provide examples in which the responses of machine learning models to changes in driving forces are very sensitive to the basins selected for training and can be implausible. On the other hand, conceptual hydrologic models generally provide an inferior fit during calibration and validation periods but seem to show a more plausible and more consistent response to changes in driving forces beyond those present in the calibration data set. However, whenever input changes are strong enough to alter catchment characteristics, such as vegetation or soil, prediction with conceptual models becomes unreliable unless the required modifications to their parameters would be known (which is generally not the case). In principle, machine learning models could be better for such predictions, as they could learn the effect of such changes in catchment characteristics from other catchments. However, as we have seen already for relatively small changes in driving forces, more research is needed to realize this potential.

Our conclusions provide a motivation for intensifying research regarding approaches that try to combine the strengths of machine learning and conceptual (or even physical) hydrologic models. Hybrid approaches that profit from physical constraints and machine learning flexibility could eliminate the problem of implausible behavior and reduce the sensitivity to the training data set of the LSTMs and, on the other hand, improve the quality of fit compared to the conceptual hydrologic models. Examples of such promising research directions are considering physical constraints or mechanisms in machine learning models (Nearing et al., 2020; Jiang et al., 2020; Xie et al., 2021), postprocessing the output of mechanistic models with machine learning models (Konapala et al., 2020), or inferring functional relationships in conceptual hydrologic models by replacing parameterized elements or functions by machine learning models (Höge et al., 2022).

https://doi.org/10.5194/hess-2023-168
Preprint. Discussion started: 26 July 2023

Appendix A: Conceptual Hydrologic Models

A1 Auxiliary Functions

425 We introduce here auxiliary functions that are used to smooth transitions between different hydrologic regimes. Smooth transitions lead to smoother posterior shapes, facilitate numerics, and are more realistic even in cases of physically sharp transitions as they represent averages over the catchment where the environmental conditions that determine the transitions are not homogeneous Kavetski et al. (2006).

We suggest two parameterizations of a smoothed Heaviside function:

430 $$f_{\mathrm{SH}}^{\mathrm{logistic}}(x,\Delta x) = \frac{1}{1+\exp\left(-4\dfrac{x}{\Delta x}\right)} \tag{A1a}$$

$$f_{\mathrm{SH}}^{\mathrm{quadratic}}(x,\Delta x) = \begin{cases} 0 & \text{for } x \le -\Delta x \\ \dfrac{1}{2}\left(\dfrac{x+\Delta x}{\Delta x}\right)^2 & \text{for } -\Delta x < x \le 0 \\ 1-\dfrac{1}{2}\left(\dfrac{\Delta x - x}{\Delta x}\right)^2 & \text{for } 0 < x \le \Delta x \\ 1 & \text{for } x > \Delta x \end{cases} \tag{A1b}$$

These functions are visualized in Figure A1. The two shapes are very similar, but note that the quadratic version is exactly zero or unity for $x \le -\Delta x$ or $x \ge \Delta x$, respectively, whereas the logistic version approaches these values asymptotically.

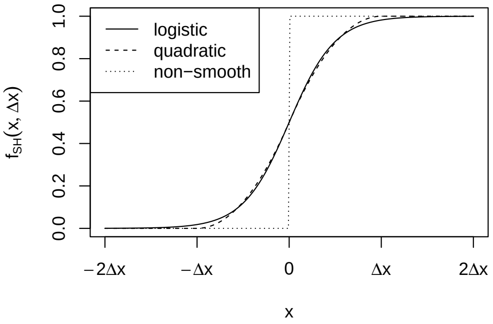

Figure A1. Shapes of the smoothed Heaviside functions.

435

The smooth transition function from zero to a linear increase is given by the equation

$$f_{\mathrm{SI}}^{\mathrm{quadratic}}(x,\Delta x) = \begin{cases} 0 & \text{for } x \leq -\Delta x \\ \dfrac{(x+\Delta x)^2}{4\Delta x} & \text{for } -\Delta x < x \leq \Delta x \\ x & \text{for } x > \Delta x \end{cases} \tag{A2}$$

and is visualized in Figure A2. Note that the function exactly matches its non-smooth version for $x \leq -\Delta x$ and for $x \geq \Delta x$.

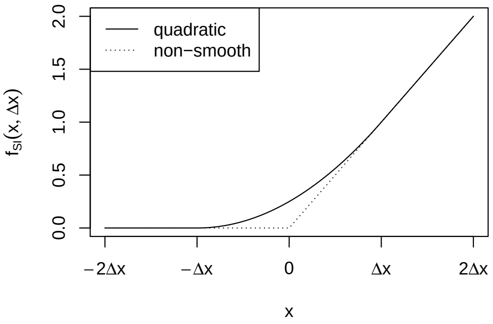

Figure A2. Shape of the smoothed start of linear increase function.

440 These two functions will be used in addition to exponential functions to formulate smooth transitions in the conceptual hydrologic models.

A2 GR4neige

The GR4J model is a conceptual hydrologic model formulated with a daily time step (J = journellement = daily) that has proven to lead to an excellent performance when only four parameters (thus the 4 in the name) are fitted for a given catchment Perrin et al. (2003). As our objective is to simulate in continuous time, we use the continuous-time version GR4 Santos et al. (2018). 445

As the set of catchments extends to high altitudes, we extend the continuous-time version of the GR4J model Santos et al. (2018) by a continuous-time version of the discrete time snow accumulation model "Cemaneige" Valery et al. (2014). We thus call this model "GR4neige" to refer to the original models. Our notation is a compromise between the original publications and the attempt to use similar parameter names across different models. Figure A3 gives a schematic overview of the model. 450

To formulate the snow model, the catchment is divided into n_b elevation bands for which precipitation and temperature input is required.

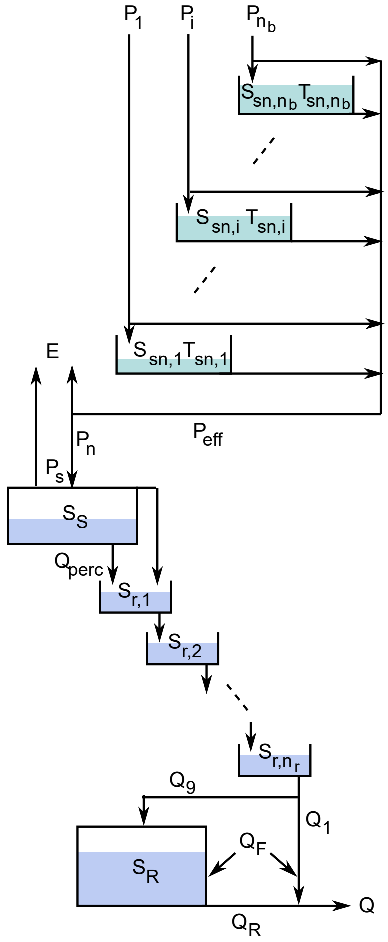

Figure A3. Schematic diagram of the GR4neige model (Santos et al., 2018; Valery et al., 2014, modified).

Precipitation is divided into snow and rain by using the fraction of rainfall calculated from daily minimum and maximum temperature as follows:

$$f_{\mathrm{snow},i} = \begin{cases} 0 & \text{for } T_{\mathrm{min},i} \geq T_{\mathrm{sf,th}} \\ \dfrac{T_{\mathrm{sf,th}} - T_{\mathrm{min},i}}{T_{\mathrm{max},i} - T_{\mathrm{min},i}} & \text{for } T_{\mathrm{max},i} > T_{\mathrm{sf,th}} \text{ and } T_{\mathrm{min},i} < T_{\mathrm{sf,th}} \\ 1 & \text{for } T_{\mathrm{max},i} \leq T_{\mathrm{sf,th}} \end{cases} . \tag{A3}$$

Here, the index i refers to the elevation band, $T_{\mathrm{min},i}$ and $T_{\mathrm{max},i}$ refer to the daily minimum and maximum temperature in the elevation band i, and $T_{\mathrm{sf,th}}$ is the threshold temperature for snowfall (see Table A1 for a list of all model parameters and their default values and ranges).

The snowpack in each elevation band is characterized by its water equivalent, $S_{\mathrm{sn},i}$, and its "cold content" indicated by a "temperature", $T_{\mathrm{sn},i}$. The function of this temperature is to delay the melting process whenever the temperature was very cold before it climbs above zero. The two state variables, $S_{\mathrm{sn},i}$ and $T_{\mathrm{sn},i}$, fulfill the following differential equations:

$$\frac{\mathrm{d}S_{\mathrm{sn},i}}{\mathrm{d}t} = f_{\mathrm{snow},i}P_i - Q_{\mathrm{melt},i} \quad , \tag{A4}$$

$$\frac{\mathrm{d}T_{\mathrm{sn},i}}{\mathrm{d}t} = \frac{-\log(\theta_{\mathrm{G2}})}{U_t}\left(T_{\mathrm{mean},i} - T_{\mathrm{sn},i}\right) \quad . \tag{A5}$$

The amount of snow (in water units) is given by a simple mass balance between accumulation and melting. Temperature follows the daily mean temperature with a rate constant of $-\log(\theta_{\mathrm{G2}})/U_t$ (note that $0 < \theta_{\mathrm{G2}} < 1$ and thus $\log(\theta_{\mathrm{G2}})$ is negative; see below for the justification of this parameterization). The snow melting rate is given by

$$Q_{\mathrm{melt},i} = \frac{\theta_{\mathrm{G1}}}{U_t} \cdot f_{\mathrm{SI}}(T_{\mathrm{mean},i} - T_{\mathrm{sm,th}}, \Delta T_{\mathrm{sm}}) \cdot f_{\mathrm{SH}}(T_{\mathrm{sn},i}, \Delta T_{\mathrm{sm}}) \cdot \left(1 - \exp\left(-\frac{S_{\mathrm{sn},i}}{S_{\mathrm{sn,th}}}\right)\right) \quad , \tag{A6}$$

which approaches the proportionality with temperature above the snow melt temperature $T_{\mathrm{mean},i} - T_{\mathrm{sm,th}}$ (f_{SI}), and if the snow temperature, $T_{\mathrm{sn},i}$, is above zero (f_{SH}), and if there still is snow present (last exponential term). These conditions are formulated by smooth transitions based on the equations (A2), (A1), and the exponential term.

Note that the analytical solution of equation (A5) under constant driving forces ($T_{\mathrm{mean},i} > T_{\mathrm{sm,th}}$) and disregarding the smoothing of the transitions is given by

$$T_{\mathrm{sn},i}(t) = T_{\mathrm{mean},i} + \left(T_{\mathrm{sn},i}(0) - T_{\mathrm{mean},i}\right)\exp\left(\frac{\log(\theta_{\mathrm{G2}})}{U_t}t\right) \quad .$$

After one day (U_t) we thus get

$$T_{\mathrm{sn},i}(U_t) = T_{\mathrm{mean},i} + \left(T_{\mathrm{sn},i}(0) - T_{\mathrm{mean},i}\right)\theta_{\mathrm{G2}} = \theta_{\mathrm{G2}}T_{\mathrm{sn},i}(0) + (1-\theta_{\mathrm{G2}})T_{\mathrm{mean},i} \quad .$$

which corresponds to the original time-discrete formulation Valery et al. (2014) and thus justifies our continuous-time approach.

Similarly to our justification for equation (A5), if $S_{\mathrm{sn},i} >> S_{\mathrm{sn,th}}$, we can neglect the exponential term in equation (A6) and if we further neglect smooting, integration over one day (U_t) leads to an integrated flux of $\theta_{\mathrm{G1}}(T_{\mathrm{mean},i} - T_{\mathrm{sm,th}})$ whereas for $S_{\mathrm{sn},i} << S_{\mathrm{sn,th}}$ we get approximately $\theta_{\mathrm{G1}}(T_{\mathrm{mean},i} - T_{\mathrm{sm,th}})S_{\mathrm{sn},i}/S_{\mathrm{sn,th}}$ as given by the discrete-time model Valery et al. (2014). This makes our model and the meaning of the parameters similar, but not identical to the Cemaneige model.

Finally, the input to the hydrologic model (per unit area) is given by the sum of the precipitation fractions falling as rain plus the sum of water from melting snow weighted by the relative areas of the elevation bands:

$$P_{\mathrm{eff}} = \sum_{i=1}^{n_{\mathrm{b}}} \frac{A_i}{A}\left((1 - f_{\mathrm{snow},i})P_i + Q_{\mathrm{melt},i}\right) \quad . \tag{A7}$$

Here, A_i is the area of the elevation band i and A is the total area of the catchment.

This continuous-time snow model is now coupled with the published continuous-time version of the GR4 model Santos et al. (2018) given by the water balance differential equations for the two reservoirs S (S_S) and R (S_R) and the cascade ($S_{\mathrm{r},i}$):

$$\frac{\mathrm{d}S_\mathrm{S}}{\mathrm{d}t} = P_\mathrm{s} - E_\mathrm{s} - Q_\mathrm{perc} \quad , \tag{A8}$$

$$\frac{\mathrm{d}S_{\mathrm{r},i}}{\mathrm{d}t} = \begin{cases} P_\mathrm{n} - P_\mathrm{s} + Q_\mathrm{perc} - \dfrac{n_\mathrm{r}}{x_4} S_{\mathrm{r},i} & \text{for } i = 1 \\ \dfrac{n_\mathrm{r}}{x_4}(S_{\mathrm{r},i-1} - S_{\mathrm{r},i}) & \text{for } i = 2, ..., n_\mathrm{r} \end{cases} \quad , \tag{A9}$$

$$\frac{\mathrm{d}S_\mathrm{R}}{\mathrm{d}t} = Q_9 + Q_\mathrm{F} - Q_\mathrm{r} \quad . \tag{A10}$$

The water fluxes in these equations are given by Santos et al. (2018)

$$P_\mathrm{n} = \begin{cases} P_\mathrm{eff} - E_\mathrm{pot} & \text{for } P_\mathrm{eff} > E_\mathrm{pot} \\ 0 & \text{for } P_\mathrm{eff} \le E_\mathrm{pot} \end{cases} \quad , \tag{A11}$$

$$E_\mathrm{n} = \begin{cases} 0 & \text{for } P_\mathrm{eff} > E_\mathrm{pot} \\ E_\mathrm{pot} - P_\mathrm{eff} & \text{for } P_\mathrm{eff} \le E_\mathrm{pot} \end{cases} \quad , \tag{A12}$$

$$P_\mathrm{s} = P_\mathrm{n}\left(1 - \left(\frac{S_\mathrm{S}}{x_1}\right)^{\alpha}\right) \quad , \tag{A13}$$

$$E_\mathrm{s} = E_\mathrm{n}\left(1 - \left(1 - \frac{S_\mathrm{S}}{x_1}\right)^{\alpha}\right) \quad . \tag{A14}$$

Note that we modified the equation

$$E_\mathrm{s} = E_\mathrm{n}\left(2\frac{S_\mathrm{S}}{x_1} - \left(\frac{S_\mathrm{S}}{x_1}\right)^{\alpha}\right)$$

$$Q_\mathrm{perc} = \frac{x_1^{1-\beta}}{(\beta - 1)U_t} \nu^{\beta - 1} S_\mathrm{S}^{\beta} \quad , \tag{A15}$$

$$Q_\mathrm{uh} = \frac{n_\mathrm{res}}{x_4} S_{\mathrm{r},n_\mathrm{res}} \quad , \tag{A16}$$

https://doi.org/10.5194/hess-2023-168
Preprint. Discussion started: 26 July 2023

$$Q_9 = \Phi Q_{\mathrm{uh}} \quad , \tag{A17}$$

$$Q_1 = (1-\Phi) Q_{\mathrm{uh}} \quad , \tag{A18}$$

$$Q_{\mathrm{F}} = \frac{x_2}{x_3^{\omega}} S_{\mathrm{R}}^{\omega} \quad , \tag{A19}$$

$$Q_{\mathrm{R}} = \frac{x_3^{1-\gamma}}{(\gamma-1)U_t} S_{\mathrm{R}}^{\gamma} \quad , \tag{A20}$$

$$Q = Q_{\mathrm{R}} + \max\bigl(0, Q_1 + Q_{\mathrm{F}}\bigr) \quad . \tag{A21}$$

Note that x_2 characterizes groundwater in- or output fed by or discharging to neighboring catchments. Set x_2 = 0 if you want to conserve mass within the catchment.

The parameters of the GR4neige model are listed together with their default values and ranges in Table A1.

A3 HBV

The HBV model is probably the most frequently used conceptual hydrologic model Bergström (1992); Lindström et al. (1997); Seibert (1999); Seibert and Vis (2012). As we use continous-time models in this paper, we develop a continuous-time model that is very similar to the original discrete-time HBV model. Figure A4 gives a schematic overview of the model.

We again distinguish n_{b} elevation bands to model snow cover. In contrast to the Cemaneige model, also the soil is resolved into these elevation bands. Within each elevation band, three state variables are used: snow, snow water (water content of the snowpack) and soil moisture.

We start with the same equation as for the GR4neige model to calculate the fraction of precipitation that falls as snow in each elevation band, i:

$$f_{\mathrm{snow},i} = \begin{cases} 0 & \text{for } T_{\min,i} \geq T_{\mathrm{sf,th}} \\ \dfrac{T_{\mathrm{sf,th}} - T_{\min,i}}{T_{\max,i} - T_{\min,i}} & \text{for } T_{\max,i} > T_{\mathrm{sf,th}} \text{ and } T_{\min,i} < T_{\mathrm{sf,th}} \\ 1 & \text{for } T_{\max,i} \leq T_{\mathrm{sf,th}} \end{cases} \quad . \tag{A22}$$

parameter	meaning	unit	default value	range*
x_1	maximum capacity of production store	mm	350	$(0,\infty)$
x_2	intercatchment exchange (inflow) coeff.	mm/d	0	$(-\infty,\infty)$
x_3	capacity parameter of routing store	mm	90	$(0,\infty)$
x_4	base time of routing cascade	d	1.7	$(0,\infty)$
θ_{G1}	maximum melting rate per degree above threshold	mm/d/°C	3	$(0,\infty)$
θ_{G2}	cold capacity delay coefficient	-	0.5	$[0,1]$
$T_{\mathrm{sf,th}}$	threshold temperature for snowfall	°C	0	$(-\infty,\infty)$
$T_{\mathrm{sm,th}}$	threshold temperature for snowmelt	°C	0	$(-\infty,\infty)$
α	production store exponent	-	2	$(1,\infty)$
β	percolation exponent	-	5	$(1,\infty)$
γ	routing store outflow exponent	-	5	$(1,\infty)$
ΔT_{sm}	temperature interval for snowmelt initiation	°C	1	
$S_{\mathrm{sn,th}}$	threshold snow level for turning off snowmelt	mm	1	
ω	intercatchment exchange exponent	-	3.5	
Φ	partion coefficient routing/outflow	-	0.9	
ν	percolation coefficient	-	4/9	
n_{b}	number of elevation bands	-	5	
n_{r}	number of routing cascade reservoirs	-	11	

Table A1. Parameters of the GR4neige model. The upper part of the table lists the parameters that are always estimated for individual catchment fits, the middle part optional parameters to be added to the set of estimated parameters, and the lower part of the table lists parameters that are kept constant for these fits. * To avoid integration problems, the ranges are more strongly constrained during optimization.

The mass balance of snow is then described by the following equation:

$$\frac{\mathrm{d}S_{\mathrm{sn},i}}{\mathrm{d}t} = c_{\mathrm{sf}} f_{\mathrm{snow},i} P_i - Q_{\mathrm{melt},i} + Q_{\mathrm{refr},i} \quad . \tag{A23}$$

Here, c_{sf} is a parameter to empirically account for errors in snow measurement and evaporation of snow. In addition to the melting flow, Q_{melt}, the HBV model considers refreezing of snow water, Q_{refr}. The melting water flow is described similarly to the GR4neige model, except that there is no cold-content or snow temperature considered:

$$Q_{\mathrm{melt},i} = c_{\mathrm{melt}} \cdot f_{\mathrm{SI}}(T_{\mathrm{mean},i} - T_{\mathrm{sm,th}}, \Delta T_{\mathrm{sm}}) \cdot \left(1 - \exp\left(-\frac{S_{\mathrm{sn},i}}{S_{\mathrm{sn,th}}}\right)\right) \quad . \tag{A24}$$

Refreezing is described similarly with the reverse temperature dependence and with a parameter c_{fr} that reduces the rate compared to melting:

$$Q_{\mathrm{refr},i} = c_{\mathrm{fr}}\, c_{\mathrm{melt}} \cdot f_{\mathrm{SI}}(T_{\mathrm{sm,th}} - T_{\mathrm{mean},i}, \Delta T_{\mathrm{sm}}) \cdot \left(1 - \exp\left(-\frac{S_{\mathrm{sw},i}}{S_{\mathrm{sw,th}}}\right)\right) \quad . \tag{A25}$$

The total water flow production in each elevation band is given by the sum of melting snow and precipitation that falls as rain, $Q_{\mathrm{melt},i} + (1 - f_{\mathrm{snow},i})P_i$. Only a fraction of this water flow feeds the snow water reservoir, as this flux is limited by the amount

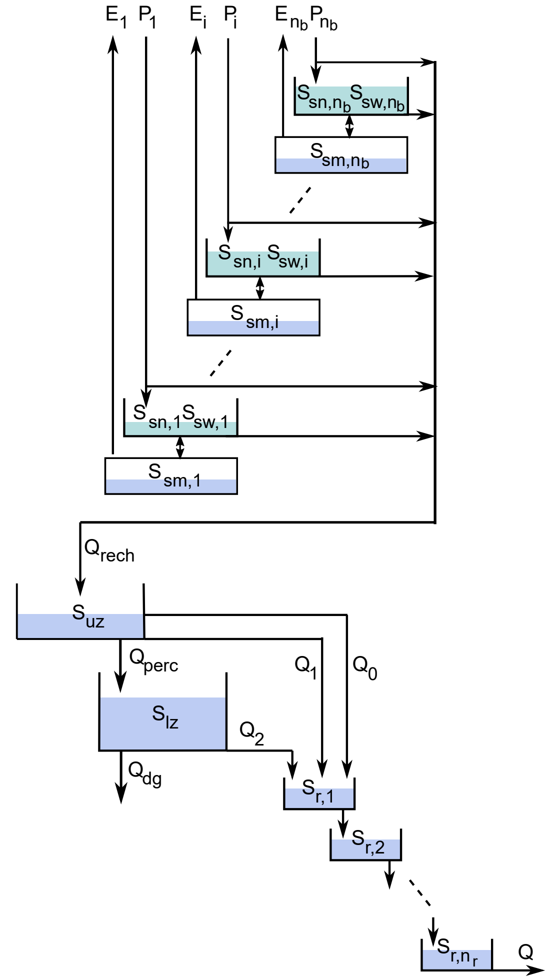

Figure A4. Schematic diagram of the HBV model as used in this paper.

of snow and by approaching the water holding capacicty of the snowpack, c_wh:

$$Q_{\mathrm{sw},i} = \begin{cases} 0 & \text{if } S_{\mathrm{sw},i} \geq c_\mathrm{wh} S_\mathrm{sn} \\ \left(Q_{\mathrm{melt},i} + (1 - f_{\mathrm{snow},i}) P_i\right) & \\ \quad \cdot \left(1 - \exp\left(-\frac{S_{\mathrm{sn},i}}{S_\mathrm{sn,th}}\right)\right) & \text{if } S_{\mathrm{sw},i} < c_\mathrm{wh} S_\mathrm{sn} \\ \quad \cdot \left(1 - \exp\left(-\frac{c_\mathrm{wh} S_{\mathrm{sn},i} - S_{\mathrm{sw},i}}{S_\mathrm{sw,th}}\right)\right) & \end{cases} \quad . \tag{A26}$$

The remaining part $Q_{\mathrm{melt},i} + (1 - f_{\mathrm{snow},i}) P_i - Q_{\mathrm{sw},i}$, together with snow water release, $Q_{\mathrm{rel},i}$, leaves the snowpack:

$$Q_{\mathrm{sn},i} = Q_{\mathrm{melt},i} + (1 - f_{\mathrm{snow},i}) P_i - Q_{\mathrm{sw},i} + Q_{\mathrm{rel},i} \quad . \tag{A27}$$

Snow water release is the most challenging part of the continuous-time formulation of the HBV model. It is needed as the relative water content would increase beyond the water colding capacity of the snowpack, c_{wh}, when snow melts even in the absence of feeding water. In the original HBV model, excess water beyond the water holding capacity is just discharged at each time step. To avoid a discontinuous flux, we accept a deviation from the discrete-time model by allowing for an increasing release of snow water already below the water holding capacity, c_{wh}:

$$Q_{\mathrm{rel},i} = \begin{cases} 0 & \text{if } S_{\mathrm{sn},i} = 0 \\ \dfrac{S_{\mathrm{sw},i}}{c_{\mathrm{wh}} S_{\mathrm{sn},i}} Q_{\mathrm{melt},i} & \text{if } S_{\mathrm{sn},i} > 0 \end{cases} \quad . \tag{A28}$$

This finally leads to the differential equation for snow water:

$$\frac{\mathrm{d}S_{\mathrm{sw},i}}{\mathrm{d}t} = Q_{\mathrm{sw},i} - Q_{\mathrm{refr},i} - Q_{\mathrm{rel},i} \quad . \tag{A29}$$

The water leaving the snowpack, $Q_{\mathrm{sn},i}$, is now divided into a fraction that feeds soil moisture and a fraction that recharges groundwater with the original nonlinear relationship with expondent β as in the original HBV model:

$$\frac{\mathrm{d}S_{\mathrm{sm},i}}{\mathrm{d}t} = Q_{\mathrm{sn},i}\left(1 - \left(\frac{S_{\mathrm{sm},i}}{S_{\mathrm{fc}}}\right)^{\beta}\right) - E_{\mathrm{pot}}\left(1 - \exp\left(-\frac{S_{\mathrm{sm}}}{S_{\mathrm{sm,th}}}\right)\right)\exp\left(-\frac{S_{\mathrm{sn}}}{S_{\mathrm{sn,th}}}\right) \quad , \tag{A30}$$

$$Q_{\mathrm{rech}} = \sum_{i=1}^{n_{\mathrm{b}}} \frac{A_i}{A} Q_{\mathrm{sn},i}\left(\frac{S_{\mathrm{sm},i}}{S_{\mathrm{fc}}}\right)^{\beta} \quad . \tag{A31}$$

Groundwater is then described by water content of an upper zone, S_{uz}, and a lower zone, S_{lz}

$$\frac{\mathrm{d}S_{\mathrm{uz}}}{\mathrm{d}t} = Q_{\mathrm{rech}} - Q_{\mathrm{perc}} - Q_0 - Q_1 \quad , \tag{A32}$$

$$\frac{\mathrm{d}S_{\mathrm{lz}}}{\mathrm{d}t} = Q_{\mathrm{perc}} - Q_2 - Q_{\mathrm{dg}} \quad , \tag{A33}$$

with a percolation flux from the upper to the lower zone given by

$$Q_{\mathrm{perc}} = c_{\mathrm{perc}}\left(1 - \exp\left(-\frac{S_{\mathrm{uz},i}}{S_{\mathrm{uz,th}}}\right)\right) \quad , \tag{A34}$$

and outfluxes given by

$$Q_0 = k_0 f_{\mathrm{SI}}(S_{\mathrm{uz}} - S_{\mathrm{uz,div}}, S_{\mathrm{uz.th}}) \quad , \tag{A35}$$

$$Q_1 = k_1 S_{\mathrm{uz}} \quad , \tag{A36}$$

$$Q_2 = k_2 S_{\mathrm{lz}} \quad , \tag{A37}$$

$$Q_{\mathrm{dg}} = k_{\mathrm{dg}} S_{\mathrm{lz}} \quad . \tag{A38}$$

These process formulations follow exactly the HBV model with the single exception of the additional flow to deep groundwater, Q_{dg}, that, if k_{dg} is negative, can also describe a feed from neighboring catchments. It turned out that some higher catchments need such a term that is similar (except from the sign) to the term characterized by the parameter x_2 of the GR4snow model.

The final model component is the a reservoir cascade that describes flow routing:

$$\frac{\mathrm{d}S_{\mathrm{r},i}}{\mathrm{d}t} = \begin{cases} Q_0 + Q_1 + Q_2 - n_{\mathrm{r}} k_{\mathrm{r}} S_{\mathrm{r},1} & \text{for } i = 1 \\ n_{\mathrm{r}} k_{\mathrm{r}} (S_{\mathrm{r},i-1} - S_{\mathrm{r},i}) & \text{for } i = 2, ..., n_{\mathrm{r}} \end{cases} \quad . \tag{A39}$$

The catchment outflow is then given as the outflow from the final routing reservoir:

$$Q = n_{\mathrm{r}} k_{\mathrm{r}} S_{\mathrm{r},n_{\mathrm{r}}} \quad . \tag{A40}$$

The parameters of this continuous-time version of the HBV model are listed together with their default values and ranges in Table A2.

parameter	meaning	unit	default value	range*
c_{melt}	maximum melting rate per degree above threshold	1/d	3	$(0,\infty)$
S_{fc}	maximum soil moisture level	mm	100	$(0,\infty)$
$S_{\mathrm{uz,div}}$	division betw. lower and upper part of upper groundw.	mm	10	$(0,\infty)$
c_{perc}	maximum percolation water flow	mm/d	1.5	$(0,\infty)$
k_0	water release coeff. from upper part of upper groundw.	1/d	1.5	$(0,\infty)$
k_1	water release coefficient from upper groundwater zone	1/d	0.1	$(0,\infty)$
k_2	water release coefficient from lower groundwater zone	1/d	0.1	$(0,\infty)$
k_r	water release coefficient of routing cascade	1/d	2	$(0,\infty)$
k_{dg}	rate coeff. for outflow from gw. to deep gw.	1/d	0	$(-\infty,\infty)$
β	exponent for water distribution to soil and groundwater	-	3	$(2,\infty)$
$T_{\mathrm{sf,th}}$	threshold temperature for snowfall	°C	0	$(-\infty,\infty)$
$T_{\mathrm{sm,th}}$	threshold temperature for snowmelt	°C	0	$(-\infty,\infty)$
ΔT_{sm}	temperature interval for snowmelt initiation	°C	1	
$S_{\mathrm{sn,th}}$	threshold snow level for turning off snowmelt	mm	1	
$S_{\mathrm{sw,th}}$	threshold snow water level for turning off refreezing	mm	0.2	
$S_{\mathrm{sm,th}}$	threshold water level of unsat. zone for turning off evap.	mm	0.5	
$S_{\mathrm{uz,th}}$	threshold for turning off percol. from upper groundw.	mm	1	
c_{fr}	coefficient of reduction of freezing rel. to melting rate	-	1	
c_{wh}	water holding fraction in snowpack	-	0.1	
c_{e}	multiplication factor for potential evaporation	-	1	
n_{b}	number of elevation bands	-	5	
n_{r}	number of routing cascade reservoirs	-	5	

Table A2. Parameters of the HBV model. The upper part of the table lists the parameters that are always estimated for individual catchment fits, the middle part optional parameters to be added to the set of estimated parameters, and the lower part of the table lists parameters that are kept constant for these fits. * To avoid integration problems, the ranges are more strongly constrained during optimization.

Appendix B: LSTM

The LSTM architecture was already successfully tested for predictions of streamflow Feng et al. (2020, 2021); Ma et al. (2021), soil moisture Fang et al. (2017, 2019, 2020), stream temperature Rahmani et al. (2021), and lake water temperature Read et al. (2019). LSTM is a type of Recurrent Neural Network (RNN) that learns from sequential data. The difference from a simple RNN is that LSTM has "memory states" and "gates", which allow it to learn how long to retain the state information, what to forget, and what to output. The forward pass of the LSTM model is described by the following equations:

$$\text{input transformation: } \mathbf{x}^t = \mathrm{ReLU}\left(\mathbf{W_I}\mathbf{I}^t + \mathbf{b_I}\right) \tag{B1}$$

$$\text{input node: } \mathbf{g}^t = \tanh\left(\mathcal{D}\left(\mathbf{W_{gx}}\mathbf{x}^t\right) + \mathbf{b_{gx}} + \mathcal{D}\left(\mathbf{W_{gh}}\mathbf{h}^{t-1}\right) + \mathbf{b_{gh}}\right) \tag{B2}$$

$$\text{input gate: } \mathbf{i}^t = \sigma\left(\mathcal{D}\left(\mathbf{W_{ix}}\mathbf{x}^t\right) + \mathbf{b_{ix}} + \mathcal{D}\left(\mathbf{W_{ih}}\mathbf{h}^{t-1}\right) + \mathbf{b_{ih}}\right) \tag{B3}$$

$$\text{forget gate: } \mathbf{f}^t = \sigma\left(\mathcal{D}\left(\mathbf{W_{fx}}\mathbf{x}^t\right) + \mathbf{b_{fx}} + \mathcal{D}\left(\mathbf{W_{fh}}\mathbf{h}^{t-1}\right) + \mathbf{b_{fh}}\right) \tag{B4}$$

$$\text{output gate: } \mathbf{o}^t = \sigma\left(\mathcal{D}\left(\mathbf{W_{ox}}\mathbf{x}^t\right) + \mathbf{b_{ox}} + \mathcal{D}\left(\mathbf{W_{oh}}\mathbf{h}^{t-1}\right) + \mathbf{b_{oh}}\right) \tag{B5}$$

$$\text{cell state: } \mathbf{s}^t = \mathbf{g}^t \odot \mathbf{i}^t + \mathbf{s}^{t-1} \odot \mathbf{f}^t \tag{B6}$$

$$\text{hidden state: } \mathbf{h}^t = \tanh\left(\mathbf{s}^t\right) \odot \mathbf{o}^t \tag{B7}$$

$$\text{output: } \mathbf{y}^t = \mathbf{W_{hy}}\mathbf{h}^t + \mathbf{b_y} \tag{B8}$$

where $\mathbf{I}^t$ represents the raw inputs for the time step, ReLU is the rectified linear unit, $\mathbf{x}^t$ is the vector to the LSTM cell, $\mathcal{D}$ is the dropout operator, $\mathbf{W}$'s are network weights, $\mathbf{b}$'s are bias parameters, σ is the sigmoidal function, $\odot$ is the element-wise multiplication operator, $\mathbf{g}^t$ is the output of the input node, $\mathbf{i}^t$, $\mathbf{f}^t$, $\mathbf{o}^t$ are the input, forget, and output gates, respectively, $\mathbf{h}^t$ represents the hidden states, $\mathbf{s}^t$ represents the memory cell states, and $\mathbf{y}^t$ is the predicted output.

The LSTM was calibrated using the catchment attributes shown in Table B1 Addor et al. (2017).

Attributes	Reduced Set	Description	Unit
elev_mean		Catchment mean elevation	m
slope_mean	x	Catchment mean slope	m/km
area_gages2	x	Catchment area (GAGESII estimate)	km^2
high_prec_freq		Frequency of high precipitation days	days/year
high_prec_dur		Average duration of high precipitation	days
low_prec_freq		Frequency of dry days	days/year
low_prec_dur		Average duration of dry periods	days
frac_forest	x	Forest fraction	-
lai_max	x	Maximum monthly mean of the leaf area index	-
lai_diff	x	Difference between max. and min. monthly mean leaf area index	-
dom_land_cover_frac	x	Fraction of catchment area associated with dominant land cover	-
dom_land_cover	x	Dominant land cover type	-
root_depth_50	x	Root depth at 50th percentiles	m
soil_depth_statsgo	x	Soil depth	m
soil_porosity	x	Volumetric soil porosity	-
soil_conductivity	x	Saturated hydraulic conductivity	cm/hr
max_water_content	x	Maximum water content	m
geol_1st_class	x	Most common geologic class in the catchment	-
geol_2nd_class	x	Second most common geologic class in the catchment	-
geol_porostiy	x	Subsurface porosity	-
geol_permeability	x	Subsurface permeability	m^2
p_mean		Mean daily precipitation	mm/day
pet_mean		Mean daily PET	mm/day
p_seasonality		Seasonality and timing of precipitation	-
frac_snow		Fraction of precipitation falling as snow	-
aridity		PET/P	-

Table B1. Full and reduced set of catchment attributes used for the calibration of the LSTM models Addor et al. (2017).

Code availability. This work is based on published and publicly available data sets and our code will be made publicly available on our institutional repository.

620 *Author contributions.* PR developed the concept of the paper, implemented the conceptual hydrologic models, did all simulations with these models and wrote the first version of the paper. KM trained the LSTMs and ran all the simulations with these models. All co-authors contributed to stimulating discussions about the paper concept and the results and to the revision of the first paper version and the finalization of the paper.

Competing interests. At least one of the (co-)authors is a member of the editorial board of *Hydrology and Earth System Sciences*.

625 *Acknowledgements.* We thank Jan Seibert for clarifications regarding the (discrete-time) HBV model. CS thanks Eawag's sabbatical support that enabled this collaboration.

References

Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: The CAMELS data set: catchment attributes and meteorology for large-sample studies, Hydrology and Earth System Sciences, 21, 5293–5313, https://doi.org/10.5194/hess-2017-169, 2017.

Alvarez-Garreton, C., Mendoza, P., Boisier, J. P., Addor, N., Galleguillos, M., Zambrano-Bigiarini, M., Lara, A., Puelma, C., Cortés, G., Garreaud, R., Mcphee, J., and Ayala, A.: The CAMELS-CL dataset: Catchment attributes and meteorology for large sample studies-Chile dataset, Hydrology and Earth System Sciences, 22, 5817–5846, https://doi.org/10.5194/hess-22-5817-2018, 2018.

Bai, P., Liu, X., and Xie, J.: Simulating runoff under changing climatic conditions: A comparison of the long short-term memory network with two conceptual hydrologic models, Journal of Hydrology, 592, 125 779, 2021.

Battjes, J. A. and Labeur, R. J.: Unsteady Flow in Open Channels, Cambridge University Press, Cambridge, UK, 2017.

Bergström, S.: The HBV Model, Tech. rep., SMHI Reports Hydrology, Sweden, 1992.

Bezanson, J., Karpinski, S., Shah, V. B., and Edelman, A.: Julia: A fast dynamic language for technical computing, arXiv preprint arXiv:1209.5145, 2012.

Bezanson, J., Edelman, A., Karpinski, S., and Shah, V. B.: Julia: A fresh approach to numerical computing, SIAM Review, 59, 65–98, 2017.

Blöschl, ü., Hall, J., Viglione, A., Perdigao, R. A. P., Parajka, J., Merz, B., Lun, D., Arheimer, B., Aronica, G. T., Bilibashi, A., Bohac, M., Bonacci, O., Borga, M., Canjevac, I., Castellarin, A., Chirico, G. B., Claps, P., Frolova, N., Ganora, D., Gorbachova, L., Gül, A., Hannaford, J., Harrigan, S., Kireeva, M., Kiss, A., Kjeldsen, T. R., Kohnova, S., Koskela, J. J., Ledvinka, O., Macdonald, N., Mavrova-Guirguinova, M., Mediero, L., Merz, R., Molnar, P., Montanari, A., Murphy, C., Osuch, M., Ovcharuk, V., Radevski, I., Salinas, J. L., Sauquet, E., Sraj, M., Szolgay, J., Volpi, E., Wilson, D., Zaimi, K., and Zivkovic, N.: Changing climate both increases and decreases European river floods, Nature, 573, 108–111, 2019.

Chagas, V. B. P., Chaffe, P. L. B., Addor, N., Fan, F. M., Fleischmann, A. S., Paiva, R. C. D., and Siqueira, V. A.: CAMELS-BR: hydrometeorological time series and landscape attributes for 897 catchments in Brazil, Earth System Science Data, 12, 2075–2096, https://doi.org/10.5194/essd-12-2075-2020, 2020.

Coxon, G., Addor, N., Bloomfield, J., Freer, J., Fry, M., Hannaford, J., Howden, N., Lane, R., Lewis, M., Robinson, E., Wagener, T., and Woods, R.: Catchment attributes and hydro-meteorological timeseries for 671 catchments across Great Britain (CAMELS-GB), https://doi.org/10.5285/8344e4f3-d2ea-44f5-8afa-86d2987543a9, 2020.

Fang, K., Shen, C., Kifer, D., and Yang, X.: Prolongation of SMAP to spatiotemporally seamless coverage of continental U.S. using a deep learning neural network, Geophysical Research Letters, 44, 11 030–11 039, 2017.

Fang, K., Shen, C., Ludwig, N., Godfrey, P., Mahjabin, T., and Doughty, C.: Combining a land surface model with groundwater model calibration to assess the impacts of groundwater pumping in a mountainous desert basin, Advances in Water Resources, 130, 12–28, 2019.

Fang, K., Kifer, D., Lawson, K., and Shen, C.: Evaluating the potential and challenges of an uncertainty quantification method for long short-term memory models for soil moisture predictions, Water Resources Research, 56, e2020WR028 095, 2020.

Feng, D., Fang, K., and Shen, C.: Enhancing Streamflow Forecast and Extracting Insights Using Long-Short Term Memory Networks With Data Integration at Continental Scales, Water Resources Research, 56, e2019WR026 793, 2020.

Feng, D., Lawson, K., and Shen, C.: Mitigating prediction error of deep learning streamflow models in large data-sparse regions with ensemble modeling and soft data, Geophysical Research Letters, 48, e2021GL092 999, 2021.

Fowler, K. J. A., Acharya, S. C., Addor, N., Chou, C., and Peel, M. C.: CAMELS-AUS: hydrometeorological time series and landscape attributes for 222 catchments in Australia, Earth Syst. Sci. Data, 13, 3847–3867, 2021.

Höge, M., Scheidegger, A., Baity-Jesi, M., Albert, C., and Fenicia, F.: Improving hydrologic models for predictions and process understanding using neural ODEs, Hydrology and Earth System Sciences, 26, 5085–5102, https://doi.org/10.5194/hess-26-5085-2022, 2022.

Hrachowitz, M., Savenije, H., Blöschl, G., McDonnell, J., Sivapalan, M., Pomeroy, J., Arheimer, B., Blume, T., Clark, M., Ehret, U., Fenicia, F., Freer, J., Gelfan, A., Gupta, H., Hughes, D., Hut, R., Montanari, A., Pande, S., Tetzlaff, D., Troch, P., Uhlenbrook, S., Wagener, T., Winsemius, H., Woods, R., Zehe, E., and Cudennec, C.: A decade of Predictions in Ungauged Basins (PUB)—a review, Hydrological Sciences Journal, 58, 1198–1255, https://doi.org/10.1080/02626667.2013.803183, 2013.

Innes, M.: Don't Unroll Adjoint: Differentiating SSA-Form Programs, CoRR, abs/1810.07951, http://arxiv.org/abs/1810.07951, 2018.

Jiang, S., Zheng, Y., and Solomatine, D.: Improving AI system awareness of geoscience knowledge: Symbiotic integration of physical approaches and deep learning, Geophys. Res. Letters, 46, e2020GL088 229, 2020.

Kavetski, D., Kuczera, G., and Franks, S. W.: Calibration of conceptual hydrological models revisited: 1. Overcoming numerical artefacts, Journal of Hydrology, 320, 173–186, 2006.

Konapala, G., Kao, S.-C., Painter, S. L., and Lu, D.: Machine learning assisted hybrid models can improve streamflow simulation in diverse catchments across the conterminous US, Environmental Research Letters, 15, 104 022, https://doi.org/10.1088/1748-9326/aba927, 2020.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Hernegger, M.: Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks, Hydrology and Earth System Sciences, 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 Resources Research, 55, 11 344–11 354, https://doi.org/10.1029/2019WR026065, 2019a.

Kratzert, F., Klotz, D., Shalev, G., Klammbauer, G., Hochreiter, S., and Nearing, G.: Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, Hydrology and Earth System Sciences, 23, 5089–5110, https://doi.org/10.5194/hess-23-5089-2019, 2019b.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., and Bergström, S.: Developmnet and test of the distributed HBV-96 hydrological model, Journal of Hydrology, 201, 272–288, 1997.

Ma, K., Feng, D., Lawson, K., Tsai, W.-P., Liang, C., Huang, X., Sharma, A., and Shen, C.: Transferring Hydrologic Data Across Continents – Leveraging Data-Rich Regions to Improve Hydrologic Prediction in Data-Sparse Regions, Water Resources Research, 57, e2020WR028 600, 2021.

Merz, R., Parajka, J., and Blöschl, G.: Time stability of catchment model parameters: Implications for climate impact analyses, Water Resources Research, 47, W02 531, 2011.

Mogensen, P. K. and Riseth, A. N.: Optim: A mathematical optimization package for Julia, Journal of Open Source Software, 3, 615, https://doi.org/10.21105/joss.00615, 2018.

Natel de Moura, C., Seibert, J., and Detzel, H. M.: Evaluating the long short-term memory (LSTM) network for discharge prediction under changing climate conditions, Hydrology Research, x, xx–xx, https://doi.org/10.2166/nh.2022.044, in press, 2022.

Nearing, G., Kratzert, F., Klotz, D., Hoedt, P.-J., Klambauer, G., Hochreiter, S., Gupta, H., Nevo, S., and Matias, Y.: A Deep Learning Architecture for Conservative Dynamical Systems: Application to Rainfall-Runoff Modeling, Workshop at NEURIPS, 2020.

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., and Gupta, H. V.: What role does hydrological science play in the age of machine learning?, Water Resources Research, 57, e2020WR028 091, 2021.

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., Viger, R., Blodgett, D., Brekke, L., Arnold, J. R., Hopson, T., and Duan, Q.: Development of a large-sample watershed-scale hydrometeorological dataset for the contiguous USA: dataset characteristics and assessment of regional variability in hydrologic model performance, Hydrology and Earth System Sciences, 19, 209–223, https://doi.org/10.5194/hess-19-209-2015, 2015.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library, in: Advances in Neural Information Processing Systems 32, pp. 8024–8035, Curran Associates, Inc., http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf, 2019.

Perrin, C., Michel, C., and Andreassian, V.: Improvement of a parsimonious model for streamflow simulation, Journal of Hydrology, 279, 275–289, 2003.

Rackauckas, C. and Nie, Q.: Differentialequations.jl–a performant and feature-rich ecosystem for solving differential equations in julia, Journal of Open Research Software, 5, 2017.

Rahmani, F., Lawson, K., Ouyang, W., Appling, A., Oliver, S., and Shen, C.: Exploring the exceptional performance of a deep learning stream temperature model and the value of streamflow data, Environmental Research Letters, 16, 024 025, 2021.

Read, J. S., Jia, X., Willard, J., Appling, A. P., Zwart, J. A., Oliver, S. K., Karpatne, A., Hansen, G. J. A., Hanson, P. C., Watkins, W., Stienbach, M., and V., K.: Process-guided deep learning predictions of lake water temperature, Water Resources Research, 55, 9173––9190, 2019.

Revels, J., Lubin, M., and Papamarkou, T.: Forward-Mode Automatic Differentiation in Julia, arXiv:1607.07892 [cs.MS], https://arxiv.org/abs/1607.07892, 2016.

Santos, L., Thirel, G., and Perrin, C.: Continuous state-space representation of a bucket-type rainfall-runoff model: a case study with the GR4 model using state-space GR4 (version 1.0), Geoscientific Model Development, 11, 1591–1605, 2018.

Seibert, J.: Regionalisation of parameters for a conceptual rainfall-funoff model, Agricultural and Forst Meteorology, 98-99, 279–293, 1999.

Seibert, J. and Vis, M. J. P.: Teaching hydrological modeling with a user-friendly catchment-runoff-model software package, Hydrology and Earth System Sciences, 16, 3315–3325, 2012.

Shen, C.: A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists, Water Resources Research, 54, 8558—8593, 2018.

Ukkola, A. M. and Prentice, I. C.: A worldwide analysis of trends in water-balance evapotranspiration, Hydrology and Earth System Sciences, 17, 4177–4187, https://doi.org/10.5194/hess-17-4177-2013, 2013.

Valery, A., Andreassian, V., and Perrin, C.: 'As simple as possible but not simpler': What is useful in a temperature-based snow-accounting routine? Part 2 – Sensitivity analysis of the Cemaneige snow accounting routine on 380 catchments, Journal of Hydrology, 517, 1176–1187, 2014.

Van Rossum, G. and Drake, F. L.: Python 3 Reference Manual, CreateSpace, Scotts Valley, CA, 2009.

Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., Chen, Y., Zeng, W., and Yu, P.: Generalizing to unseen domains: A survey on domain generalization, IEEE Transactions on Knowledge and Data Engineering, 2022.

Xie, K., Liu, P., Zhang, J., Han, D., Wang, G., and Shen, C.: Physics-guided deep learning for rainfall-runoff modeling by considering extreme events and monotonic relationships, Journal of Hydrology, 603, 127 043, 2021.

Xie, X., Ho, J. W. K., Murphy, C., Kaiser, G., Xu, B., and Chen, T. Y.: Testing and validating machine learning classifiers by metamorphic testing, The Journal of Systems and Software, 84, 544–558, 2011.

Yang, Y. and Chui, T. F. M.: Reliability assessment of machine learning models in hydrological predictions through metamorphic testing, Water Resources Research, 57, e2020WR029 471, 2021.