Hydrol. Earth Syst. Sci., 28, 2505–2529, 2024
https://doi.org/10.5194/hess-28-2505-2024

Hydrology and Earth System Sciences
Open Access

Metamorphic testing of machine learning and conceptual hydrologic models

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

[1]Eawag: Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland
[2]Institute of International Rivers and Eco-Security, Yunnan University, Kunming, China
[3]Yunnan Key Laboratory of International Rivers and Transboundary Eco-security, Yunnan University, Kunming, China
[4]Civil and Environmental Engineering, Pennsylvania State University, University Park, State College, PA, USA
[☆]retired
[★]These authors contributed equally to this work.

Correspondence: Peter Reichert (peter.reichert@emeriti.eawag.ch)

Received: 10 July 2023 – Discussion started: 26 July 2023
Revised: 15 April 2024 – Accepted: 20 April 2024 – Published: 13 June 2024

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 prediction requires a model, but the impossibility of testing such predictions against observed data makes it 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 testing 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 analyze differences in predictions between different model structures and calibration options in addition to the qualitative test of the expectations. In our case study, we apply this approach to selected conceptual and machine learning hydrological models calibrated for 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 and in addition to 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 (Kratzert et al., 2018; Shen, 2018; Kratzert et al., 2019a, b; Razavi, 2021; Ng et al., 2023; Feng et al., 2020). In particular, it has been shown that the training of machine learning models jointly with 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:

1. using machine learning models, in particular deep learning architectures in the form of long short-term memory

(LSTM) models, that are highly flexible and contain a large number of parameters;

2. training the models jointly on large sets of diverse catchments 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 (Kratzert et al., 2018, 2019a, b). It has been shown that this kind of hydrologic knowledge can even be transferred across continents (Ma et al., 2021). The success demonstrated by a large number of studies based on machine learning models trained on 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 the strong support of hydrologic expert knowledge (Hrachowitz et al., 2013; Nearing et al., 2021). The availability of many more data sets for other countries besides the USA, such as Chile (Alvarez-Garreton et al., 2018), Great Britain (Coxon et al., 2020), Brazil (Chagas et al., 2020), Australia (Fowler et al., 2021), Switzerland (Höge et al., 2023), and more, bears great potential for further development of hydrologic modeling across catchments, continents, and climatic regions.

The primary focus of the studies cited above was on model training and validation for a future part of the time series or for catchments not used for calibration. The question of whether this success is transferable to the prediction of the consequences of modified driving forces in these catchments has been less investigated (Bai et al., 2021; Natel de Moura et al., 2022; Wi and Steinschneider, 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 are no data available for validating the models under such perturbations. The problem is of a different nature for conceptual hydrologic models compared to 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 structure, etc. (Merz et al., 2011). Furthermore, it is also difficult to extend the models to mechanistically describe these changes. The prediction with machine learning models could lead to wrong results due to a 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 catchments used for training. For these reasons, it is of interest to compare the predictions of both kinds of models under modified driving forces and to investigate whether the results depend on the training data set or 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 the deficiencies of models and pathways for their improvement based on these results. Such attempts have been made before and have uncovered problems in the predictions of LSTM models (Bai et al., 2021; Razavi, 2021; Natel de Moura et al., 2022; Wi and Steinschneider, 2022). We extend this kind of study by considering precipitation changes in addition to temperature changes (this has been done in some of the previous studies); by using LSTM models trained on a large set of catchments (this has been done in some of the previous studies); by investigating responses for different elevation classes separately to reduce the uncertainty in the response predicted by the experts; and by including sensitivity analyses regarding catchment attributes, basins used for calibration, and the numerical seed of the optimization algorithm. We will do model simulations with isolated changes in precipitation and temperature and compare the resulting change in outlet discharge with the expected outcomes for selected basins from the CAMELS data set (Newman et al., 2015; Addor et al., 2017). Note that this is a metamorphic testing design (Xie et al., 2011; Yang and Chui, 2021) that facilitates the formulation of the expected qualitative behavior rather than a realistic climate change scenario that would consist of coupled temperature and precipitation changes with a more complex time dependence. 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

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

2. 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 test and complements 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 outputs while still allowing 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 patterns, as would be needed for real predictions. Such additional tests to model fit are important as it has been shown that the quality of fit and prediction accuracy do not necessarily improve in parallel. At least one case study came to the following 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 judgments 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 use a multi-model approach and extend the test to the analysis of the sensitivity of the results to model structure and to different training or calibration options. In particular, we compare conceptual and machine learning approaches as we expect complementary strengths and weaknesses. Conceptual approaches, due to their consideration of (simplified) physical principles, can be expected to provide reliable predictions if the input changes are small enough to not considerably alter catchment properties, such as vegetation and soil structure. On the other hand, machine learning models may be more critical for out-of-sample predictions, but due to the high diversity of catchments used for training, they bear the potential to also consider changes in catchment properties. This allows us to identify the quantitative deviations of predictions (in relation to modified inputs) between model structures. The investigation of the sensitivity of the predictions to calibration options further provides insight into the robustness of the results of the metamorphic test. The chosen model structures are described in more detail in Sect. 2.2 and in Appendix A and B, and the complementary calibration options are presented in Sect. 3.3.

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 structures and increases the trust in the reliability of the predictions.

b. *Metamorphic test succeeded, but quantitative responses of different models disagree.* The predicted response is in qualitative agreement with the expectations, but the quantitative response is sensitive to the investigated model structures or to aspects of the model calibration process. This result shows the limits of metamorphic testing, but the identified differences between responses may still stimulate reflections on model structure improvements.

c. *Metamorphic test failed, some models inconsistent with others.* The predicted response is sensitive to the investigated model structures or aspects of the model calibration process, with some responses being in agreement and others being in disagreement with the expert expectations. This indicates problems in some models reliably predicting the response to the investigated input changes and indicates the need for a revision of model structures or training processes.

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 caused either by similar deficits in all model structures that lead to wrong predictions or by incomplete scientific knowledge that leads to incorrect expert predictions. This is the most difficult outcome of the metamorphic analysis, but it still demonstrates the importance of the analysis as it uncovers 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 similar structural deficits in all investigated models. This may initiate an extended research process that depends on the investigated system and models.

For metamorphic testing, we choose simple, isolated changes in precipitation (increase by 10 %) and temperature (increase by 1°) to make it easier for experts to characterize the expected response. As mentioned before, this setup covers inputs relevant for climate change predictions, but it does not represent realistic input changes for climate change. Figure 1 shows a visualization of the simplified expected response of the catchment outlet discharge to these changes; this is discussed in more detail below. The simplified expected responses shown in Fig. 1 represent general trends;

the true expected response will be less smooth due to shorter-term precipitation and temperature fluctuations.

1. The first input change is a constant relative increase in precipitation by 10 %. The investigated response we are interested is in the change in discharge at the catchment outlet resulting from the change in precipitation:

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

   where Q is the hydrologic model describing catchment outlet discharge as a function of the precipitation time series, P, and the 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. Note that, according to Eq. (1), such a relative change does not lead to any input change during periods without precipitation. An alternative absolute input modification would not make sense for precipitation as this would lead to the elimination of dry weather periods. In terms of the expected response, as shown in the top row of Fig. 1, 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 traveling of the flood wave do 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 structure, and other catchment properties. For more complex and stronger input changes, more complex response patterns are possible, as discussed by Blöschl et al. (2019). In a worldwide 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 Fig. 5 in Ukkola and Prentice, 2013). However, as this is an analysis of past data, many other factors also changed; 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. A second input change is a constant increase in temperature by 1 °C. The investigated response we are interested in is the change in discharge at the catchment outlet resulting from the change in temperature:

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

   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 Eq. (1). In terms of the expected response, as shown in the bottom-left panel of Fig. 1, 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 traveling 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 a seasonal discharge pattern dominated by snow cover dynamics, we expect an increase in river discharge in autumn or winter due to a later change in precipitation from rain to snowfall and an earlier melting in spring followed by a decrease in river discharge because the snowmelt will be complete earlier. This response pattern is shown in the bottom-right panel of Fig. 1. 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 Fig. 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.

As the training data contained precipitation- or temperature-related catchment attributes, such as mean daily precipitation and the fraction of precipitation falling as snow, we compared the results to training with an omission of these kinds of attributes to avoid biased results due to inconsistent changes in driving forces. Table B1 in Appendix B lists the full sets, as well as the reduced sets, of catchment attributes used for this comparison.

The intention of our study is to identify potential problems in hydrologic models and to learn from them and not to provide a representative overview of the 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 in improving models on which we do not focus in this paper, we only select catchments for which all our primary modeling approaches lead to a very good fit during the calibration period (Nash–Sutcliffe efficiency (NSE) > 0.8 during the calibration period for all investigated model structures; the range of NSE values for the selected catchments was 0.82–0.92). All of these models also lead to a good fit during the validation period (range of NSE values of 0.67–0.91). To best represent the conditions for which we can describe the expected response as described above, we choose the following:

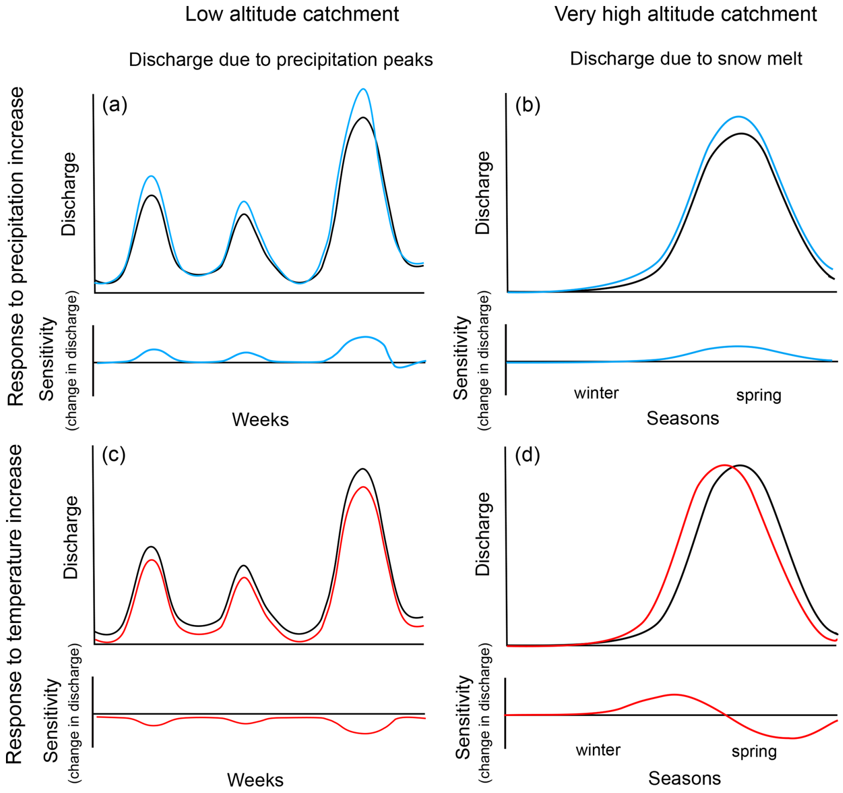

Figure 1. Simplified expected response to a precipitation increase of 10 % **(a, b)** and to a temperature increase of 1° **(c, d)** for low-altitude catchments (**a, c**; response to precipitation events within weeks) and for high-altitude catchments (**b, d**; changes in seasonal snowmelt peak). Black lines: discharge for unmodified input. Blue lines: discharge and sensitivity (change in discharge) for modified precipitation input. Red lines: discharge and sensitivity (change in discharge) for modified temperature input.

- *Low-altitude warm basins.* These basins should only have a minor amount of snow and thus a relatively simple response pattern, as described above.

- *Very-high-altitude (cold) basins.* In these basins, the response should be dominated by the shifts in snowfall and snowmelt.

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 and 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.

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) and 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 the 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 in terms of their consideration of the basins for calibration and in terms of the set of catchment attributes used for calibration (see Sect. 2.3 below). Appendix B provides an overview of the setup of these models.

2.3 Calibration and training

The parameter values of the models were obtained by optimization of a loss function that quantifies the deviation of model output from observations, as described below. As it corresponds to the typical use in the literature, for this optimization, we use the term calibration for the conceptual models, and we use the term training for the machine learning models.

The conceptual hydrologic models used daymet altitude band inputs as provided in the CAMELS data (https://ral.ucar.edu/solutions/products/camels, last access: 29 January 2021, Newman et al., 2014), aggregated to a maximum of five bands, for catchment-by-catchment calibration by maximization of the posterior with a simple, uncorrelated, normal-error model and wide priors. Optimization was performed using the LBFGS algorithm (Liu and Nocedal, 1989). As there are only incomplete banded input data available for the basin nos. 12167000, 12186000, and 12189500, we calibrated the model for only 668 of the 671 basins of the US CAMELS data set.

The LSTM model was jointly trained for all 671 basins of the US CAMELS data set (https://ral.ucar.edu/solutions/products/camels, last access: 29 January 2021) 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). Optimization was performed using the AdaDelta optimizer with the parameters lr = 1.0 and rho = 0.9 (Zeiler, 2012). As we encountered some unexpected responses in the low-altitude basins to a change in temperature (see Sect. 3.2.1 below), additional training was done, as described in Sect. 3.3.

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) (1 October 1980–30 September 19950 for calibration and 1 October 1995–30 September 2010 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), and Optim (Mogensen and Riseth, 2018).

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

All our code is publicly available (conceptual models: https://doi.org/10.25678/000CQ0, Reichert et al., 2024, LSTM: https://doi.org/10.5281/zenodo.3993880, Shen, 2020).

3 Results and discussion

3.1 Quality of fit

Figure 2 provides an overview of the Nash–Sutcliffe Efficiency (NSE) values achieved for the calibration and validation periods for all modeling approaches in the 668 basins for which the conceptual models could also be calibrated, as well as in 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 the calibration and validation periods. The LSTM has an additional advantage in that it generalizes very well to catchments not used for training, but this feature is not investigated in this paper.

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 Sect. 2. For each of the three classes, we selected four basins. Figure 3 provides an overview of the locations of the selected four basins within each category. As mentioned in Sect. 2.1, these basins were selected by allowing for an excellent fit for all modeling approaches (NSE > 0.8 during the calibration period; the range of NSE values across models and selected catchments was 0.82–0.92 for the calibration period and 0.67–0.91 for the validation period). Due to the limited number of basins in these categories, the strong requirement regarding the quality of fit for all modeling approaches, and the wish to have the same number of basins in each category, it was not possible to compare more basins. However, as shown in the following sections, there are notably consistent patterns in the responses to changes within each of these categories.

3.2.1 Low-altitude warm basins

Figure 4 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 Figs. S2 to S17 in the Supplement. These results are systematic across all studied basins, demonstrating that the features discussed in this section represent the typical behavior of this kind of basin and are not just an artifact of the specific basin and year.

As is shown by the NSE values in the legends of the fourth panels (Figs. 4 and S2 to S17) for these basins, all of the compared primary modeling 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 are larger – mostly much larger – than 0.65 during validation; see also the overview of NSE values in Fig. 2).

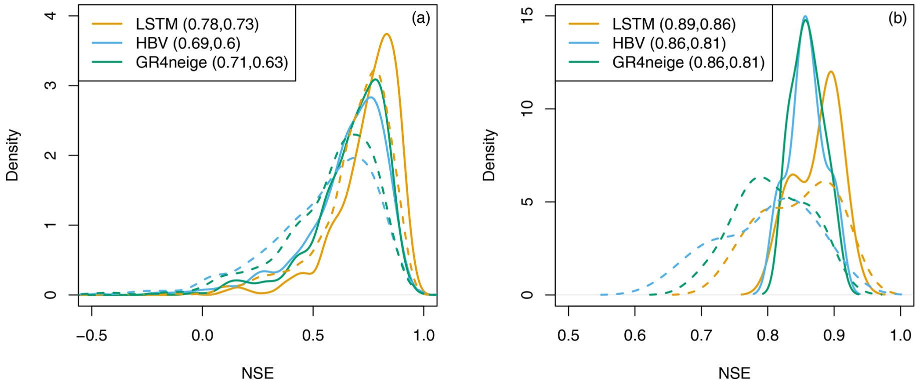

Figure 2. Overview of NSE values of all modeling approaches for the calibration period (solid) and the validation period (dashed) for all 668 basins **(a)** and for the 12 basins selected for metamorphic testing **(b)**; note the different scale of the x axis). The median NSE values are indicated in brackets in the legend (calibration period, validation period).

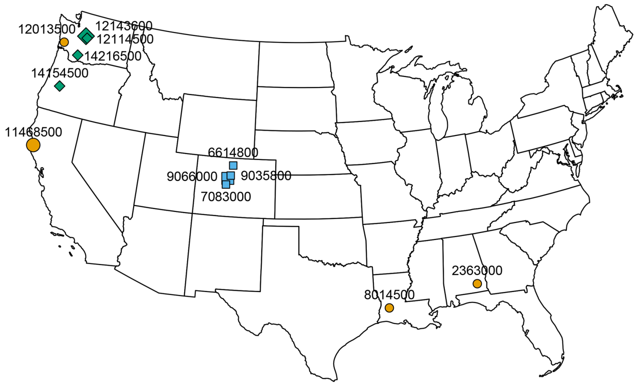

Figure 3. Basins used for metamorphic testing. Orange circles indicate low-altitude basins, blue squares indicate very-high-altitude basins, and green rhombs indicate intermediate-altitude basins. The large markers represent basins with results shown in the main paper; the results for all basins are shown in the Supplement. The numbers represent CAMELS basin identifiers. Map produced with the R package `usmap`: https://CRAN.R-project.org/package=usmap (last access: 10 March 2024).

The sensitivities to a 10 % increase in precipitation, ΔQ_P (see Eq. 1), are plotted in the top panel of Fig. 4 (and of Figs. S2 to S17). All our modeling approaches (GR4neige, HBV, LSTM) lead to very similar sensitivities to the investigated relative change in precipitation. The sensitivities to the investigated increase in precipitation also correspond to our expectations, as described in Sect. 2.1 (see, in particular, Fig. 1, top-left panel), 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 category A outlined in Sect. 2.1 (consistent agreement with expectations across modeling approaches) and makes us confident in the response of all models to changes in precipitation.

In contrast to the precipitation sensitivities shown in the first panel, the second panel of Fig. 4 (and of Figs. S2 to S17) shows substantial differences in temperature sensitivities, ΔQ_T (see Eq. 2), between different modeling approaches (GR4neige, HBV, LSTM). The sensitivities of the hydrologic models GR4neige and HBV are essentially negative (the discharge for increased temperature is smaller than it was with the original temperature), with only some brief positive excursions associated with small shifts in discharge peaks. These are the expected sensitivities as described in Sect. 2.1 (see, in particular, Fig. 1, bottom-left panel). 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 Sect. 2.1 (inconsistency with expectations for some model structures). This raises the question of which approach may provide the correct response. The conceptual hydrologic models may share similar deficits with the expected response as both are based on similar expert knowledge. On the other hand, the LSTM may, due to its broad coverage of the climatic conditions of 671 CAMELS basins, better consider the effect of changing catchment properties resulting from the increasing temperature, or its response may be incorrect due to poor out-of-sample prediction. Since we see here a striking difference in the behavior 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 Sect. 3.3.

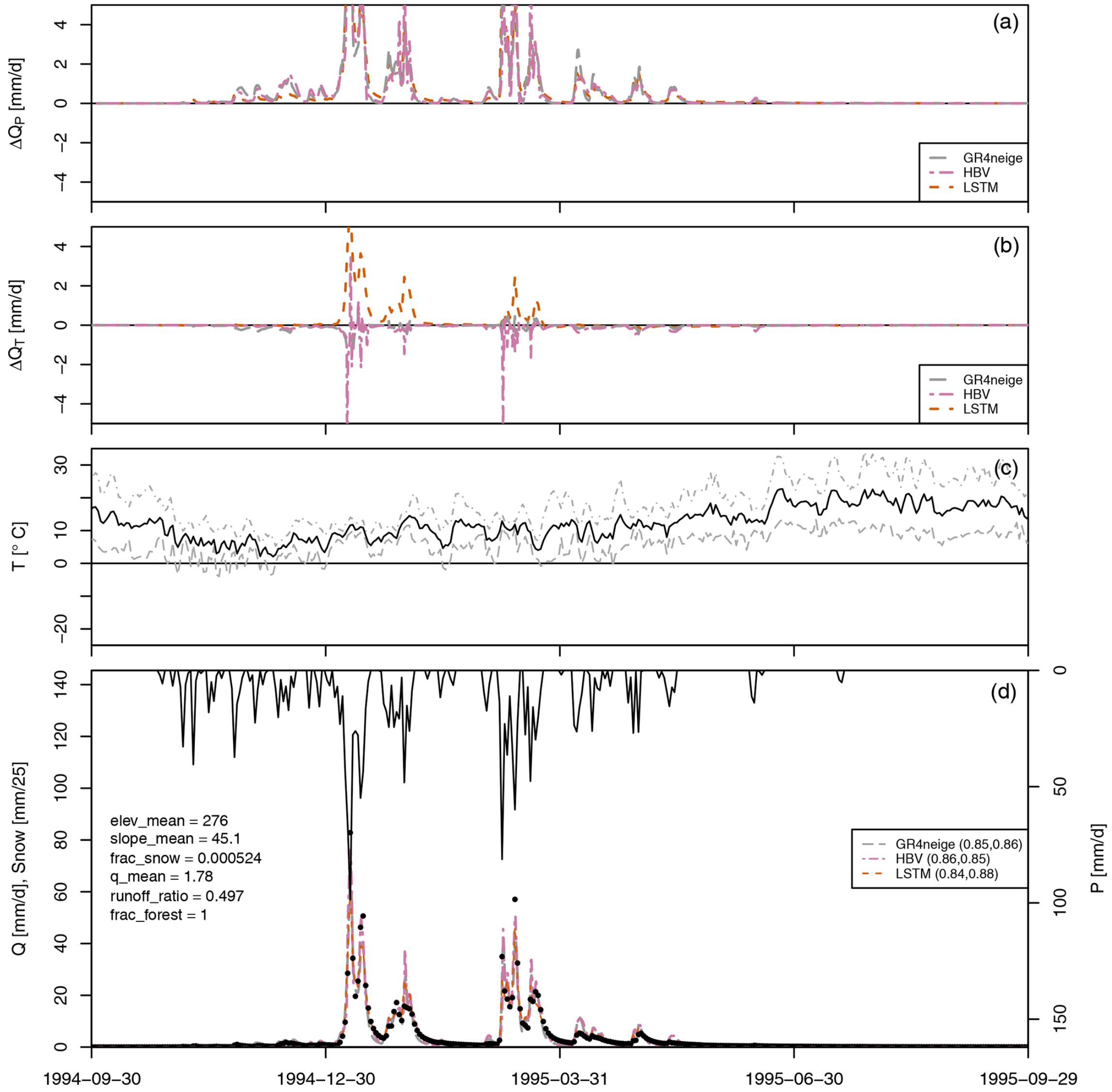

Figure 4. Results for basin no. 11468500. (**a**) Modeled sensitivity of discharge to a 10 % increase in precipitation, ΔQ_P (see Eq. 1). (**b**) Modeled sensitivity of discharge to a 1° increase in temperature, ΔQ_T (see Eq. 2). (**c**): Minimum, mean, and maximum temperature. (**d**): Observed precipitation (from top, right axis); modeled (lines) and observed (circles) discharge and modeled snow cover in max. five altitude bands (dashed lines, 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).

3.2.2 Very-high-altitude (cold) basins

Figure 5 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 Figs. S18 to S33 in the Supplement. These results demonstrate that the features discussed in this section represent the typical behavior of this kind of basins and are not just an artifact of the specific basin and year.

The legends of the fourth panels in these figures (Figs. 5 and S18 to S33) show again that we have an excellent fit during the calibration and validation periods for all modeling approaches (GR4neige, HBV, LSTM), with NSE values larger than 0.8 during the calibration periods and larger than 0.7 during the validation periods.

The precipitation sensitivities show, in this case, more differences than for the low-altitude catchments in Sect. 3.2.1. All models show the expected positive precipitation sensitivities (higher discharge for higher precipitation), but the response of the LSTM is considerably smaller and smoother than the responses of the conceptual models. Still, these results correspond qualitatively to our expectations, as described in Sect. 2.1 (see, in particular, Fig. 1, top-right panel).

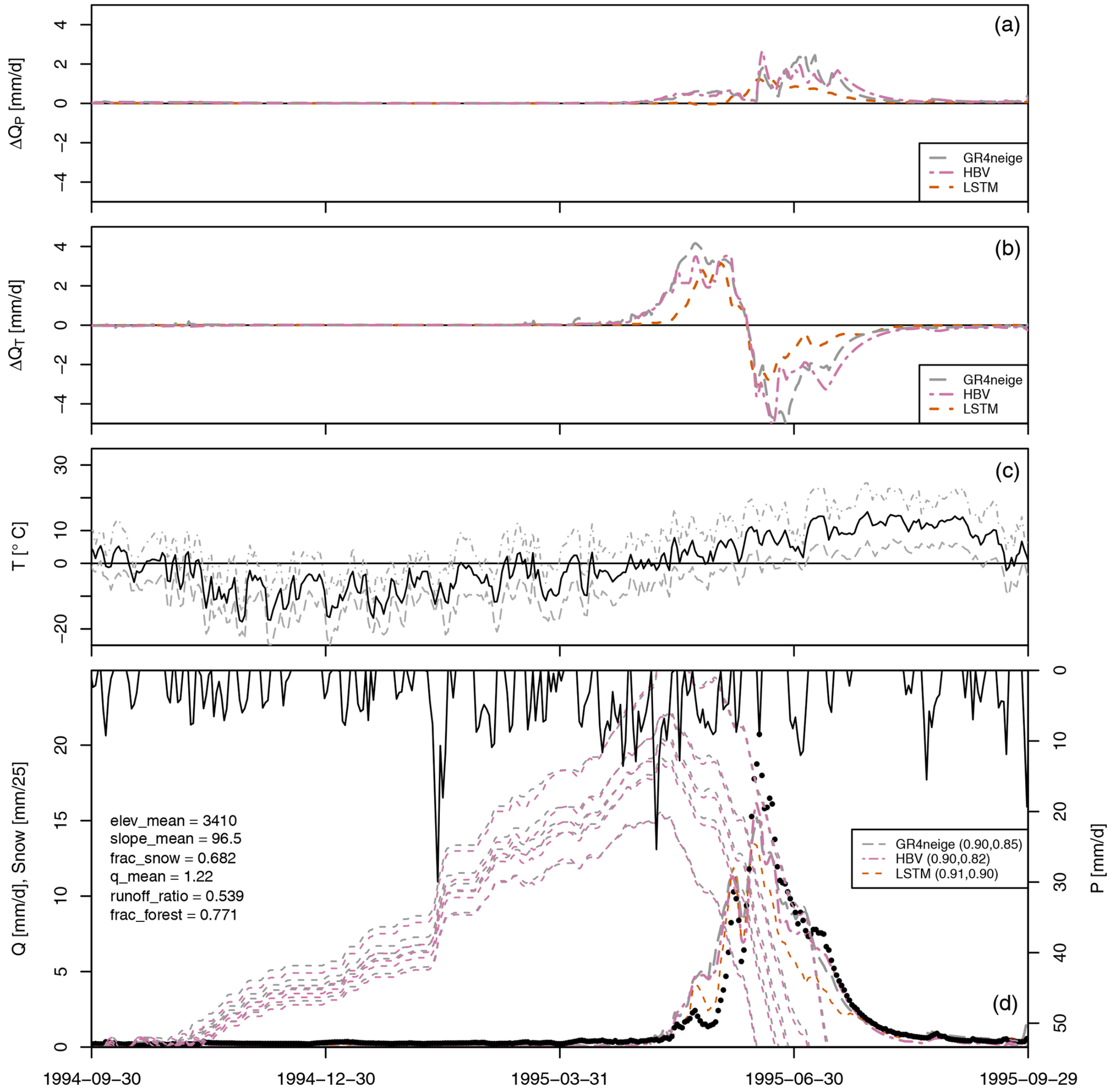

Figure 5. Results for basin no. 09066000. **(a)** Modeled sensitivity of discharge to a 10 % increase in precipitation, ΔQ_P (see Eq. 1). **(b)** Modeled sensitivity of discharge to a 1° increase in temperature, ΔQ_T (see Eq. 2). **(c)** Minimum, mean, and maximum temperature. **(d)** Observed precipitation (from top, right axis); modeled (lines) and observed (circles) discharge and modeled snow cover in max. five altitude bands (dashed lines, left axis); 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).

Also, the temperature sensitivities show the expected behavior of a positive sensitivity (higher discharge for higher temperature) due to the earlier snowmelt process followed by a negative sensitivity (lower discharge for higher temperature) due to the earlier completion of the snowmelt process; see Sect. 2.1, in particular Fig. 1, bottom-right panel). There is a tendency where the positive response starts later and the negative response ends earlier in the LSTM model compared to the conceptual hydrologic models. Also, these sensitivities tend to be smaller and smoother for the LSTM model than for the conceptual hydrologic models.

The results of these metamorphic tests thus belong to category B outlined in Sect. 2.1 (significant differences between approaches but still in qualitative agreement with the expectations). As mentioned in Sect. 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 the calibration and validation periods can still differ considerably in their response to modified driving forces. This indicates that one should be cautious with the predictions of such responses.

3.2.3 Intermediate-altitude basins

Figure 6 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 Figs. S34 to S49 in the Supplement. These results demonstrate that the features discussed in this section represent the typical behavior of these kinds of basins and are not just an artifact of the specific basin and year.

The legends of the fourth panels in these figures (Figs. 6 and S34 to S49) show again that we have an excellent fit during the calibration and validation periods for all modeling approaches (GR4neige, HBV, LSTM), with NSE values larger than 0.8 during the calibration periods and larger than 0.77 during the validation periods.

The results shown in Fig. 6 combine the results discussed in the previous sections but resemble more closely the high-altitude catchments as snow cover still dominates the dynamic behavior during most of the season.

Figure 6a and b 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 with respect 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 model. In this part of the season, the responses of the LSTM model are smoother and smaller than those of the hydrologic models. Again, the qualitative nature of metamorphic testing makes it difficult to assess which of these results are more plausible. These results are again in category B of our result classification for metamorphic testing, as outlined in Sect. 2.1.

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 model 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
- the 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 model 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 is much less relevant but could be used as a proxy for temperature by the LSTM model) and removing attributes that would have to be modified for prediction with modified inputs (e.g., all precipitation-related attributes, such as mean daily precipitation, 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 to 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 were caused by a convergence into a "bad" minimum while other local minima would have lead to better results. Table 1 lists the model versions used for this sensitivity analysis.

Figure 7 shows the precipitation and temperature sensitivities of these models at a higher scale and for a shorter time period than in Fig. 4 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 Figs. S50 to S65 in the Supplement. The results are qualitatively similar throughout all catchments and periods. The precipitation sensitivities, ΔQ_P (see Eq. 1), are quite insensitive to any of these modifications from the original setup (see upper panel in Fig. 7 and in Figs. S50 to S65). In particular, it is remarkable that omitting the mean precipitation (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 do not change the observed precipitation sensitivities.

The results for the temperature sensitivities, ΔQ_T (see Eq. 2), are quite different for the different modeling approaches (see lower panel in Fig. 7 and in Figs. S50 to S65). In particular, those of the original LSTM model and LSTM_515 are mostly positive, whereas those of LSTM_211 (trained only with the basins with mean elevations smaller than 300 m) are negative, except for a single peak after the tick mark for 19 March 1995. The sensitivities of LSTM_361 are less visible due to the congestion in the figure, but the trend from the original LSTM model to LSTM_211 is very clear.

To better visualize the differences between the temperature sensitivities of all modeling approaches, Fig. 8 shows a quantification of these differences. As in metamorphic testing, the quantitative change in results is difficult to assess, and so we focus on the quantification of differences in sensitivities for which the signs are different between different model versions. We therefore calculated the mean-squared differences in the sensitivities of all the combinations of approaches, setting the differences to zero if the sensitivities have the same signs and omitting periods of very low sensitivities ($< 0.1\,\mathrm{mm\,d^{-1}}$). Of these values, we took the

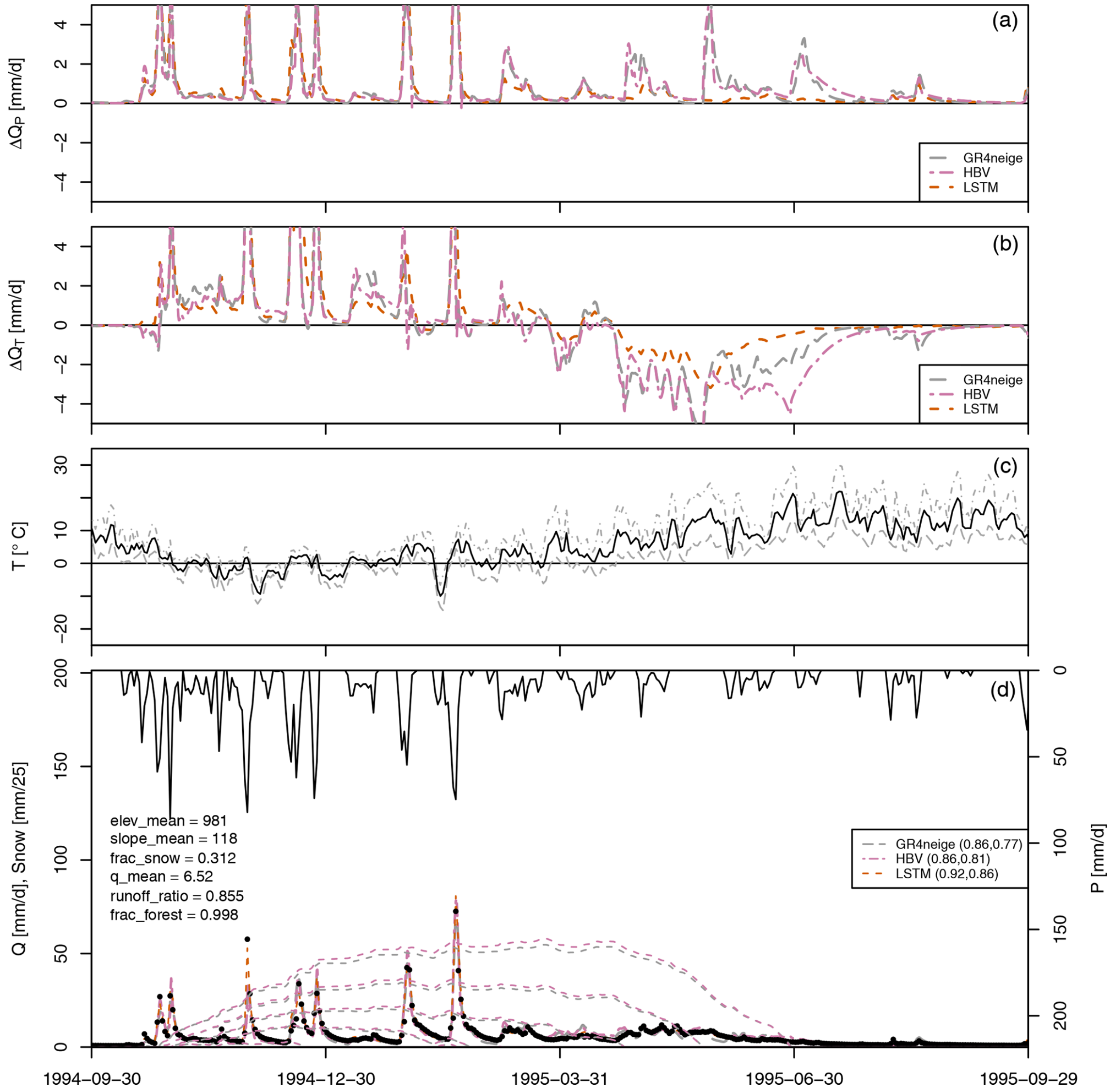

Figure 6. Results for basin no. 12143600. **(a)** Modeled sensitivity of discharge to a 10 % increase in precipitation, ΔQ_P (see Eq. 1). **(b)** Modeled sensitivity of discharge to a 1° increase in temperature, ΔQ_T (see Eq. 2). **(c)** Minimum, mean, and maximum temperature. **(d)** Observed precipitation (from top, right axis); modeled (lines) and observed (circles) discharge and modeled snow cover in max. five altitude bands (dashed lines, left axis); 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).

Table 1. Overview of models. The first three rows describe the basic models planned for use 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 model for low-altitude catchments.

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

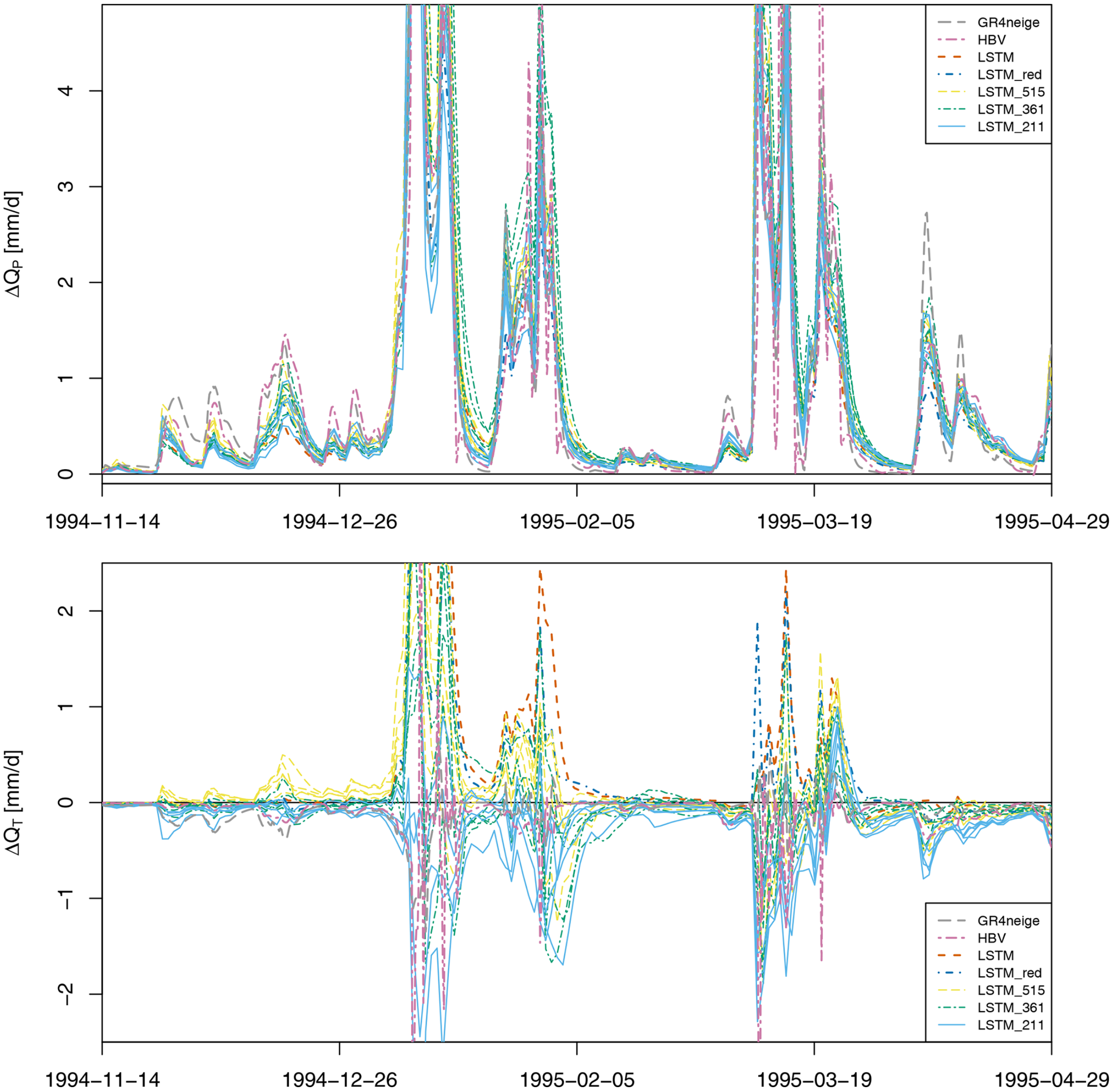

Figure 7. Sensitivities for basin no. 11468500 and different LSTM calibration options (see Table 1). The sensitivities for the GR4neige, HBV, and LSTM models are the same as those shown in Fig. 4. In addition, LSTM_red shows the sensitivities when calibrating with a reduced set of catchment attributes, 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 five different seed values are shown. See text for more details.

square root and the mean across all basins of the same class (here, low-altitude catchments) and across different random-number seeds. These results are shown in Fig. 8.

The results in the first two columns clearly show that the sensitivities of the LSTM model are considerably different from those of the GR4neige and HBV models and that they approach the results of these models when moving from the original LSTM model over LSTM_red, LSTM_515, and LSTM_361 to LSTM_211 (see Table 1 for model definitions). In parallel, in this order, the sensitivities deviate more and more from those of the original LSTM model (third column in Fig. 8). These quantitative difference measures thus clearly confirm the qualitative discussion in the previous paragraph.

The same quantification as shown for the temperature sensitivities in Fig. 8 was performed for precipitation and temperature sensitivities and is shown in Fig. S1 in the Supplement. These results clearly demonstrate that there is no similar problem in the signs of calculated sensitivities for precipitation for all classes of basins and for temperature for the very-high-elevation basins. The temperature sensitivities for the intermediate-elevation basins also show problems between the modeling approaches, but these are more difficult to interpret as they combine effects for low- and high-altitude catchments.

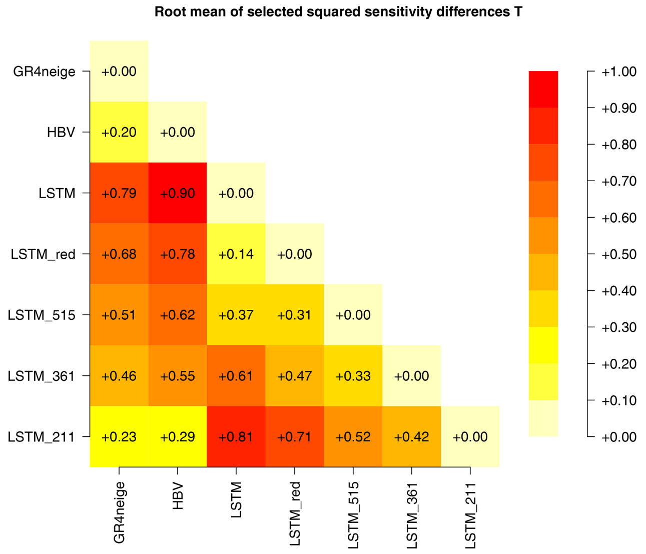

Figure 8. Differences between the temperature sensitivities, ΔQ_T, of all modeling approaches, quantified as described in the text ($\mathrm{mm\,d^{-1}}$).

Table 2. Summary of results of metamorphic testing. See Sect. 2.1 for a description of the result categories (note that the lowest three models were designed for low-altitude catchments and were therefore not used for intermediate- and high-altitude catchments). The most problematic outcomes (C) are marked in bold.

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**	B	B	B	B
LSTM_515	A	**C**				
LSTM_361	A	**C**				
LSTM_211	A	A				

Table 2 summarizes the results of the extended metamorphic testing.

Our sensitivity analysis demonstrates that, for the LSTM applied to the low-altitude catchments, there is a large sensitivity to the training set of modeled catchment outlet responses resulting from input temperature changes. It is particularly remarkable that the sensitivities to temperature change are different despite all of the different calibration versions providing an excellent fit during the calibration and validation periods. This suggests that the problem in 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 the calibration and validation and for catchments not used for calibration, as 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 indicate that the LSTM trained with all basins is not as universal as expected from previous results but that an LSTM trained more specifically with the low-altitude basins passes the metamorphic test. It is possible that pre-conditioning the model to these catchments could have a similar effect.

3.4 Discussion

Our results uncovered problems in LSTM models in producing consistent (primarily across training data) results for predictions related to input changes beyond those used during training. These results provide a motivation for investigating approaches that either (A) widen the diversity of basins used for training to expand the potential for learning physical mechanisms (Wi and Steinschneider, 2022) or (B) try to combine the strengths of machine learning and conceptual (or even physical) hydrologic models (Ng et al., 2023; Shen et al., 2023; Tsai et al., 2021; Jiang et al., 2020). As approach A failed for the low-level basins in our study (when extending the training data set to high-elevation basins), our results indicate that approach B seems more promising for systems that we have a good mechanistic knowledge of. Examples of such hybrid approaches are considering physical constraints or mechanisms in machine learning models (Nearing et al., 2020; Razavi, 2021; Xie et al., 2021; Zhong et al., 2023), post-processing the output of mechanistic models with machine learning models (Konapala et al., 2020), using conceptual model outputs or components (such as evapotranspiration estimates) as additional input to machine learning models (Wi and Steinschneider, 2022), or inferring functional relationships in conceptual hydrologic models by replacing parameterized elements or functions with machine learning models (Jiang et al., 2020; Tsai et al., 2021; Höge et al., 2022). In the last category, which some referred to as differentiable modeling (Shen et al., 2023), neural networks are seamlessly connected to programmatically differentiable (permitting gradient tracking) process-based equations, and they are trained together in an end-to-end fashion (Jiang et al., 2020; Feng et al., 2022; Bindas et al., 2024). This framework may address the sensitivity problem by hard-coding (and thus guaranteeing) required physical sensitivities to forcings and attributes as prior equations or restricting information flows into and out of neural networks, which should be investigated in the future. If sensitivity is a primary concern, one should also use caution with neural networks as post-processing layers as they can modify the assumed sensitivities.

These results thus lead to the following conclusions regarding the questions raised in the introduction:

- A good fit during calibration and validation periods does not guarantee a good response to changes in driving forces. There is a strong need to analyze 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.

- With regard to the usefulness of metamorphic testing, this is a very useful tool to test models beyond the usual calibration–validation process. The problem with metamorphic testing is that it requires the response to be at least qualitatively known. This can be difficult and even biased as this requires inputs of expert knowledge that can be biased by the current state of incomplete scientific knowledge. For this reason, we strongly recommend using "extended metamorphic testing", in which we not only check model predictions for modified inputs with the expected response but, as in a multi-model approach, 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 number of "B" classifications listed in 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 are not directly relevant to a specific prediction (in our case, to 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 provide information directly relevant to the question to be investigated – improves the quality of fit and the response to input changes. However, it may raise awareness of the necessity for carefully selecting training data as adding less relevant data (for this specific question) may have adverse effects.

- With regard to machine learning vs. conceptual models, the modeling 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, they need to be calibrated based on a large set of basins, and their performance can be poor when calibrated based on a single basin. In addition, 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 need much less data; in particular, they can easily be applied to a single basin. They 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 properties, such as vegetation or soil structure, prediction with conceptual models becomes unreliable unless the required modifications to their parameters are known or the vegetation and soil structure are part of the model. The latter would lead to a model with mechanisms that are very difficult to parameterize with sufficient accuracy. In principle, machine learning models could be better for such predictions as they could learn the effect of such changes in catchment properties from other catchments. However, as we have seen already for relatively small changes in driving forces, more research is needed to realize this potential. This would require the development of models that are trained on a training data set that contains a large diversity of catchments with different characteristics and/or that are constrained or pre-conditioned by physical and biological considerations.

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 US CAMELS data set, for which all modeling 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 generally provide 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 the main aim of this study.

- In an extended 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 being that 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 of low-altitude basins, for which the LSTM models 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, which should represent the factors with a direct physical influence, did not resolve this issue. However, training the LSTM on only low-elevation catchments reversed the sign of the sensitivities, which then mostly agreed with those of the conceptual hydrologic models. As for the intermediate- and high-elevation basins, the response of the LSTM model was then in qualitative agreement with that of the conceptual models but was generally smaller and smoother.

Our results indicate the need for caution in the prediction of LSTM models for inputs that were not present in a similar form in the training data set. Enlarging the training set to situations that are not of direct relevance to the investigated problem may even deteriorate the results. In our case study, this occurred with results for low-altitude basins when also including high-altitude basins for training. On the other hand, enlarging the set of low-altitude basins improved the response for low-altitude basins, which is in agreement with the experience with machine learning models that a large training set of basins is important for leveraging their full potential. These results 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 of the LSTM models to the training data set and, on the other hand, improve the quality of fit compared to the conceptual hydrologic models.

Appendix A: Conceptual hydrologic models

A1 Auxiliary functions

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).

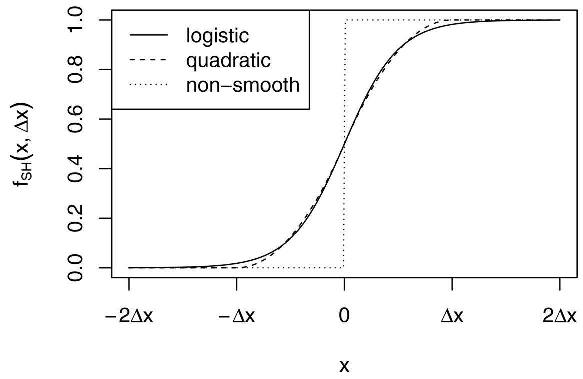

Figure A1. Shapes of the smoothed Heaviside functions.

We suggest two parameterizations of a smoothed Heaviside function:

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

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

These functions are visualized in Fig. 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.

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

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

and this is visualized in Fig. A2.

Note that the function exactly matches its non-smooth version for $x \le -\Delta x$ and for $x \ge \Delta x$.

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

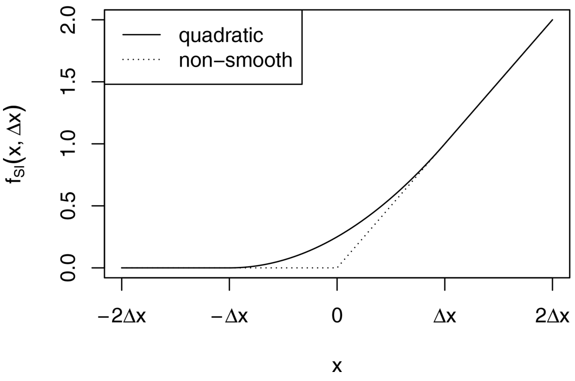

Figure A2. Shape of the smoothed start of the linear-increase function.

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).

As the set of catchments extends to high altitudes, we extend the continuous-time version of the GR4J model (Santos et al., 2018) with 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.

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

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

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

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 is 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:

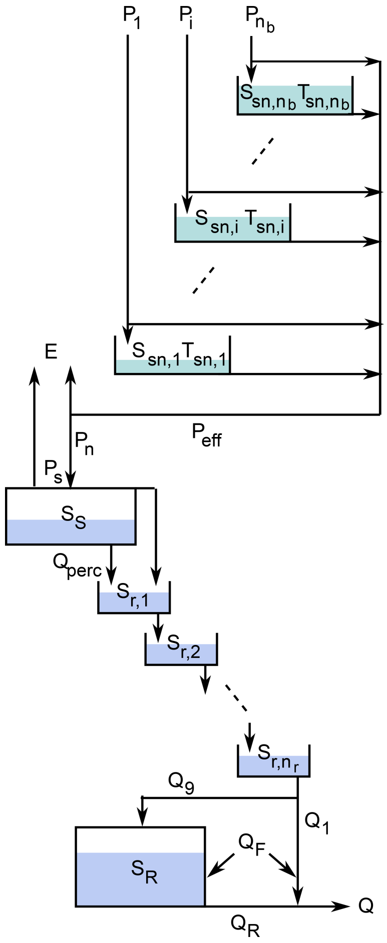

Figure A3. Schematic diagram of the GR4neige model (Santos et al., 2018; Valery et al., 2014, modified).

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

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

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

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

which approaches the proportionality with temperature above the snowmelt 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 Eqs. (A2) and (A1b) and the exponential term.

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

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

After 1 d (U_t), we thus get

$$\begin{aligned} T_{\mathrm{sn},i}(U_t) &= T_{\mathrm{mean},i} + \left(T_{\mathrm{sn},i}(0) - T_{\mathrm{mean},i}\right)\theta_{\mathrm{G2}} \\ &= \theta_{\mathrm{G2}} T_{\mathrm{sn},i}(0) + (1-\theta_{\mathrm{G2}}) T_{\mathrm{mean},i}, \end{aligned}$$

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

Similarly to our justification for Eq. (A5), if $S_{\mathrm{sn},i} \gg S_{\mathrm{sn,th}}$, we can neglect the exponential term in Eq. (A6), and if we further neglect smoothing, integration over 1 d (U_t) leads to an integrated flux of $\theta_{\mathrm{G1}}(T_{\mathrm{mean},i} - T_{\mathrm{sm,th}})$, whereas for $S_{\mathrm{sn},i} \ll S_{\mathrm{sn,th}}$, 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:

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

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 ($S_{\mathrm{r},i}$):

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

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

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

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

$$P_{\mathrm{n}} = \left\{ \begin{array}{ll} P_{\mathrm{eff}} - E_{\mathrm{pot}} & \text{for } P_{\mathrm{eff}} > E_{\mathrm{pot}} \\ 0 & \text{for } P_{\mathrm{eff}} \leq E_{\mathrm{pot}} \end{array} \right., \tag{A11}$$

$$E_{\mathrm{n}} = \left\{ \begin{array}{ll} 0 & \text{for } P_{\mathrm{eff}} > E_{\mathrm{pot}} \\ E_{\mathrm{pot}} - P_{\mathrm{eff}} & \text{for } P_{\mathrm{eff}} \leq E_{\mathrm{pot}} \end{array} \right., \tag{A12}$$

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

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

Note that we modified the equation:

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

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

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

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

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

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

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

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

Note that x_2 characterizes groundwater input or output fed by or discharging into 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 continuous-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, the soil is also 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:

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 lists 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.

Parameter	Meaning	Unit	Default value	Range*
x_1	Maximum capacity of production store	mm	350	$(0, \infty)$
x_2	Intercatchment exchange (inflow) coeff.	$\mathrm{mm\,d^{-1}}$	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	$\mathrm{mm\,d^{-1}\,°C^{-1}}$	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	Portion of 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	

* To avoid integration problems, the ranges are more strongly constrained during optimization.

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

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

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

Here, c_{sf} is a parameter to empirically account for errors in snow measurement and the 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 in the GR4neige model, except that there is no cold content or snow temperature considered:

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

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

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

The total water flow production in each elevation band is given by the sum of melting snow and precipitation that fall 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 of snow and by approaching the water-holding capacity of the snowpack, c_{wh}:

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

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:

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

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-holding 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}:

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 lists 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.

Parameter	Meaning	Unit	Default value	Range*
c_{melt}	Maximum melting rate per degree above threshold	mm (°C d^{-1})	3	$(0, \infty)$
S_{fc}	Maximum soil moisture level	mm	100	$(0, \infty)$
$S_{\mathrm{uz,div}}$	Division between lower and upper parts of upper groundw.	mm	10	$(0, \infty)$
c_{perc}	Maximum percolation water flow	mm d^{-1}	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_{sf}	Snowfall correction factor	–	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	

* To avoid integration problems, the ranges are more strongly constrained during optimization.

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

This finally leads to the differential equation for snow water:

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

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 the exponent β, as in the original HBV model:

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

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

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

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

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

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

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

while outfluxes are given by

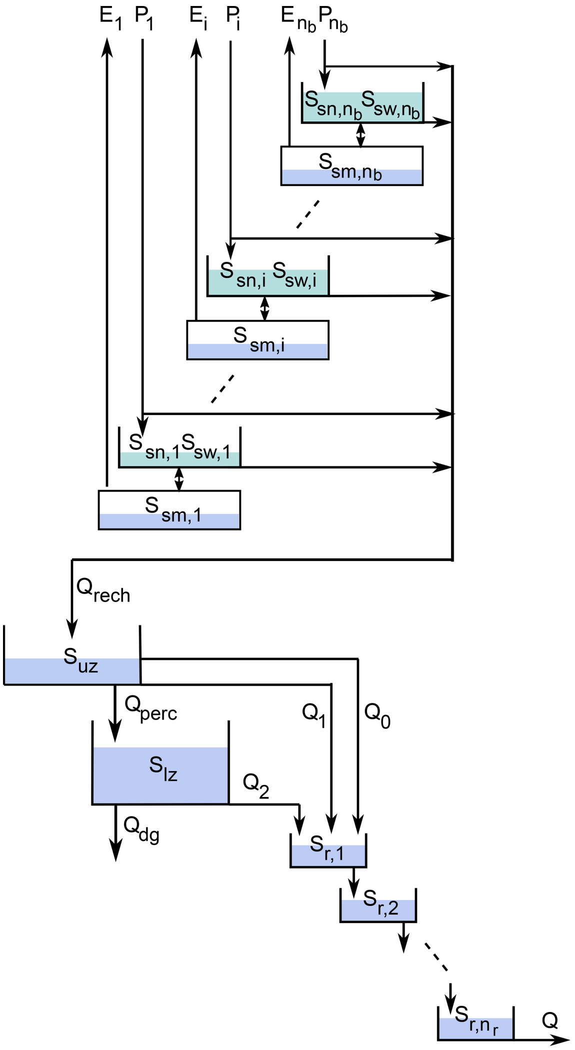

Figure A4. Schematic diagram of the HBV model as used in this paper.

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

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

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

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

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 for 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:

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

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

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

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

Appendix B: LSTM

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

Attributes	Reduced set	Description	Unit
elev_mean		Catchment mean elevation	m
slope_mean	x	Catchment mean slope	$\mathrm{m\,km^{-1}}$
area_gages2	x	Catchment area (GAGESII estimate)	$\mathrm{km^2}$
high_prec_freq		Frequency of high-precipitation days	$\mathrm{d\,yr^{-1}}$
high_prec_dur		Average duration of high precipitation	days
low_prec_freq		Frequency of dry days	$\mathrm{d\ yr^{-1}}$
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	$\mathrm{cm\,h^{-1}}$
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	$\mathrm{m^2}$
p_mean		Mean daily precipitation	$\mathrm{mm\,d^{-1}}$
pet_mean		Mean daily PET	$\mathrm{mm\,d^{-1}}$
p_seasonality		Seasonality and timing of precipitation	–
frac_snow		Fraction of precipitation falling as snow	–
aridity		PET/P	–

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), snow water equivalent (Song et al., 2024; Cui et al., 2023), lake water temperature (Read et al., 2019), dissolved oxygen (Zhi et al., 2023), and nitrate (Saha et al., 2023). 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 equations outlined below.

$$\text{Input transformation: } \boldsymbol{x}^t = \mathrm{ReLU}\big(\mathbf{W}_I \boldsymbol{I}^t + \boldsymbol{b}_I\big) \tag{B1}$$

$$\text{Input node: } \boldsymbol{g}^t = \tanh\Big(\mathcal{D}\big(\mathbf{W}_{gx}\boldsymbol{x}^t\big) + \boldsymbol{b}_{gx} + \mathcal{D}\big(\mathbf{W}_{gh}\boldsymbol{h}^{t-1}\big) + \boldsymbol{b}_{gh}\Big) \tag{B2}$$

$$\text{Input gate: } \boldsymbol{i}^t = \sigma\Big(\mathcal{D}\big(\mathbf{W}_{ix}\boldsymbol{x}^t\big) + \boldsymbol{b}_{ix} + \mathcal{D}\big(\mathbf{W}_{ih}\boldsymbol{h}^{t-1}\big) + \boldsymbol{b}_{ih}\Big) \tag{B3}$$

$$\text{Forget gate: } \boldsymbol{f}^t = \sigma\Big(\mathcal{D}\big(\mathbf{W}_{fx}\boldsymbol{x}^t\big) + \boldsymbol{b}_{fx} + \mathcal{D}\big(\mathbf{W}_{fh}\boldsymbol{h}^{t-1}\big) + \boldsymbol{b}_{fh}\Big) \tag{B4}$$

$$\text{Output gate: } \boldsymbol{o}^t = \sigma\Big(\mathcal{D}\big(\mathbf{W}_{ox}\boldsymbol{x}^t\big) + \boldsymbol{b}_{ox} + \mathcal{D}\big(\mathbf{W}_{oh}\boldsymbol{h}^{t-1}\big) + \boldsymbol{b}_{oh}\Big) \tag{B5}$$

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

$$\text{Hidden state: } \boldsymbol{h}^t = \tanh\big(\boldsymbol{s}^t\big) \odot \boldsymbol{o}^t \tag{B7}$$

$$\text{Output: } y^t = \mathbf{W}_{hy}\boldsymbol{h}^t + \boldsymbol{b}_y \tag{B8}$$

In the above, $\boldsymbol{I}^t$ represents the raw inputs for the time step; ReLU is the rectified linear unit; $\boldsymbol{x}^t$ is the vector to the LSTM cell; $\mathcal{D}$ is the dropout operator; $\mathbf{W}$ refers to network weights; $\boldsymbol{b}$ refers to bias parameters; σ is the sigmoidal function; $\odot$ is the element-wise multiplication operator; $\boldsymbol{g}^t$ is the output of the input node; $\boldsymbol{i}^t$, $\boldsymbol{f}^t$, and $\boldsymbol{o}^t$ are the input, forget, and output gates, respectively; $\boldsymbol{h}^t$ represents the hidden states; $\boldsymbol{s}^t$ represents the memory cell states; and $\boldsymbol{y}^t$ is the predicted output.

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

Code and data availability. This work is based on published and publicly available data sets (https://doi.org/10.5065/D6MW2F4D, Newman et al., 2021), and our code is publicly available (conceptual models: https://doi.org/10.25678/000CQ0, Reichert et al., 2024, LSTM: https://doi.org/10.5281/zenodo.3993880, Shen, 2020).

Supplement. The supplement related to this article is available online at: https://doi.org/10.5194/hess-28-2505-2024-supplement.

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 LSTM models and ran all the simulations with these models. All the co-authors contributed to the 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*. The peer-review process was guided by an independent editor, and the authors also have no other competing interests to declare.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors.

Acknowledgements. We thank Jan Seibert for the clarifications regarding the (discrete-time) HBV model. Chaopeng Shen thanks Eawag's sabbatical support that enabled this collaboration. Finally, we thank Scott Steinschneider and two anonymous reviewers for their comments and suggestions that helped improve the paper.

Review statement. This paper was edited by Yongping Wei and reviewed by two anonymous referees.

References

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

Alvarez-Garreton, C., Mendoza, P. A., Boisier, J. P., Addor, N., Galleguillos, M., Zambrano-Bigiarini, M., Lara, A., Puelma, C., Cortes, G., Garreaud, R., McPhee, J., and Ayala, A.: The CAMELS-CL dataset: catchment attributes and meteorology for large sample studies – Chile dataset, Hydrol. Earth Syst. Sci., 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, J. Hydrol., 592, 125779, https://doi.org/10.1016/j.jhydrol.2020.125779, 2021.

Battjes, J. A. and Labeur, R. J.: Unsteady Flow in Open Channels, Cambridge University Press, Cambridge, UK, ISBN 978-1-107-15029-4, 2017.

Bergström, S.: The HBV Model, Tech. rep., SMHI Reports Hydrology, Sweden, https://www.smhi.se/polopoly_fs/1.83589!/Menu/general/extGroup/attachmentColHold/mainCol1/file/RH_4.pdf (last access: 20 January 2022), 1992.

Bezanson, J., Karpinski, S., Shah, V. B., and Edelman, A.: Julia: A fast dynamic language for technical computing, arXiv [preprint], https://doi.org/10.48550/arXiv.1209.5145, 2012.

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

Bindas, T., Tsai, W.-P., Liu, J., Rahmani, F., Feng, D., Bian, Y., Lawson, K., and Shen, C.: Improving River Routing Using a Differentiable Muskingum-Cunge Model and Physics-Informed Machine Learning, Water Resour. Res., 60, e2023WR035337, https://doi.org/10.1029/2023WR035337, 2024.

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 Syst. Sci. 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), NERC Environmental Information Data Centre, https://doi.org/10.5285/8344e4f3-d2ea-44f5-8afa-86d2987543a9, 2020.

Cui, G., Anderson, M., and Bales, R.: Mapping of snow water equivalent by a deep-learning model assimilating snow observations, J. Hydrol., 616, 128835, https://doi.org/10.1016/j.jhydrol.2022.128835, 2023.

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, Geophys. Res. Lett., 44, 11030–11039, 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, Adv. Water Resour., 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 Resour. Res., 56, e2020WR028095, https://doi.org/10.1029/2020WR028095, 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 Resour. Res., 56, e2019WR026793, https://doi.org/10.1029/2019WR026793, 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, Geophys. Res. Lett., 48, e2021GL092999, https://doi.org/10.1029/2021GL092999, 2021.

Feng, D., Liu, J., Lawson, K., and Shen, C.: Differentiable, learnable, regionalized process-based models with multiphysical outputs can approach state-of-the-art hydrologic prediction accuracy, Water Resour. Res., 58, e2022WR032404, https://doi.org/10.1029/2022WR032404, 2022.

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, https://doi.org/10.5194/essd-13-3847-2021, 2021.

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

Höge, M., Kauzlaric, M., Siber, R., Schönenberger, U., Horton, P., Schwanbeck, J., Floriancic, M. G., Viviroli, D., Wilhelm, S., Sikorska-Senoner, A. E., Addor, N., Brunner, M., Pool, S., Zappa, M., and Fenicia, F.: CAMELS-CH: hydro-meteorological time series and landscape attributes for 331 catchments in hydrologic Switzerland, Earth Syst. Sci. Data, 15, 5755–5784, https://doi.org/10.5194/essd-15-5755-2023, 2023.

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, Hydrolog. Sci. J., 58, 1198–1255, https://doi.org/10.1080/02626667.2013.803183, 2013.

Jiang, S., Zheng, Y., and Solomatine, D.: Improving AI system awareness of geoscience knowledge: Symbiotic integration of physical approaches and deep learning, Geophys. Res. Lett., 46, e2020GL088229, https://doi.org/10.1029/2020GL088229, 2020.

Kavetski, D., Kuczera, G., and Franks, S. W.: Calibration of conceptual hydrological models revisited: 1. Overcoming numerical artefacts, J. Hydrol., 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, Environ. Res. Lett., 15, 104022, https://doi.org/10.1088/1748-9326/aba927, 2020.

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

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., and Nearing, G. S.: Toward improved predictions in ungauged basins: Exploiting the power of machine learning, Water Resour. Res., 55, 11344–11354, https://doi.org/10.1029/2019WR026065, 2019a.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G.: Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, Hydrol. Earth Syst. Sci., 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, J. Hydrol., 201, 272–288, 1997.

Liu, D. C. and Nocedal, J.: On the limited memory BFGS method for large scale optimization, Math. Program., 45, 503–528, 1989.

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 Resour. Res., 57, e2020WR028600, https://doi.org/10.1029/2020WR028600, 2021.

Merz, R., Parajka, J., and Blöschl, G.: Time stability of catchment model parameters: Implications for climate impact analyses, Water Resour. Res., 47, W02531, https://doi.org/10.1029/2010WR009505, 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, Hydrol. Res., 53, 657–667, https://doi.org/10.2166/nh.2022.044, 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, Virtual Workshop AI for Earth Sciences, NeurIPS, 12 December 2020, https://ai4earthscience.github.io/neurips-2020-workshop/papers/ai4earth_neurips_2020_51.pdf (last access: 1 November 2021), 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 Resour. Res., 57, e2020WR028091, https://doi.org/10.1029/2020WR028091, 2021.

Newman, A., Sampson, K., Clark, M. P., Bock, A., Viger, R. J., and Blodgett, D.: A large-sample watershed-scale hydrometeorological dataset for the contiguous USA, Boulder, CO, UCAR/NCAR [data set], https://doi.org/10.5065/D6MW2F4D, 2014.

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

Ng, K. W., Huang, Y. F., Koo, C. H., Chong, K. L., El-Shafie, A., and Ahmed, A. N.: A review of hybrid deep learning applications for streamflow forecasting, J. Hydrol., 625, 130141, https://doi.org/10.1016/j.jhydrol.2023.130141, 2023.

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, Curran Associates, Inc., Adv. Neur. In., 32, 8024–8035, http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf (last access: 6 June 2024), 2019.

Perrin, C., Michel, C., and Andreassian, V.: Improvement of a parsimonious model for streamflow simulation, J. Hydrol., 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, 15, https://doi.org/10.5334/jors.151, 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, Environ. Res. Lett., 16, 024025, https://doi.org/10.1088/1748-9326/abd501, 2021.

Razavi, S.: Deep Learning explained: Fundamentals, explainability, and bridgeability to process-based modelling, Environ. Modell. Softw., 144, 105159, https://doi.org/10.1016/j.envsoft.2021.105159, 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 Kumar, V.: Process-guided deep learning predictions of lake water temperature, Water Resour. Res., 55, 9173–9190, https://doi.org/10.1029/2019WR024922, 2019.

Reichert, P., Ma, K., Hoge, M., Fenicia, F., Baity-Jesi, M., Feng, D., and Shen, C.: Data for: Metamorphic Testing of Machine Learning and Conceptual Hydrologic Models, Eawag: Swiss Federal Institute of Aquatic Science and Technology [code], https://doi.org/10.25678/000CQ0, 2024.

Revels, J., Lubin, M., and Papamarkou, T.: Forward-Mode Automatic Differentiation in Julia, arXiv [preprint], https://doi.org/10.48550/arXiv.1607.07892, 2016.

Saha, G. K., Rahmani, F., Shen, C., Li, L., and Cibin, R.: A deep learning-based novel approach to generate continuous daily stream nitrate concentration for nitrate data-sparse watersheds, Sci. Total Environ., 878, 162930, https://doi.org/10.1016/j.scitotenv.2023.162930, 2023.

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), Geosci. Model Dev., 11, 1591–1605, https://doi.org/10.5194/gmd-11-1591-2018, 2018.

Seibert, J.: Regionalisation of parameters for a conceptual rainfall-funoff model, Agr. Forest Meteorol., 98–99, 279–293, 1999.

Seibert, J. and Vis, M. J. P.: Teaching hydrological modeling with a user-friendly catchment-runoff-model software package, Hydrol. Earth Syst. Sci., 16, 3315–3325, https://doi.org/10.5194/hess-16-3315-2012, 2012.

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

Shen, C.: MHPI-hydroDL, Zenodo [code], https://doi.org/10.5281/zenodo.3993880, 2020.

Shen, C., Appling, A. P., Gentine, P., Bandai, T., Gupta, H., Tartakovsky, A., Baity-Jesi, M., Fenicia, F., Kifer, D., Li, L., Liu, X., Ren, W., Zheng, Y., Harman, C. J., Clark, M., Farthing, M., Feng, D., Kumar, P., Aboelyazeed, D., RahmaniHylke, F., Beck, E., Bindas, T., Dwivedi, D., Fang, K., Höge, M., Rackauckas, C., Roy. T., Xu, C., and Lawson, K.: Differentiable modelling to unify machine learning and physical models for geosciences, Nat. Rev. Earth Environ., 4, 552–567, 2023.

Song, Y., Tsai, W.-P., Gluck, J., Rhoades, A., Zarzycki, C., McCrary, R., Lawson, K., and Shen, C.: LSTM-based data integration to improve snow water equivalent prediction and diagnose error sources, J. Hydrometeorol., 25, 223–237, 2024.

Tsai, W.-P., Feng, D., Pan, M., Beck, H., Lawson, K., Yang, Y., Liu, J., and Shen, C.: From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling, Nat. Commun., 12, 5988, https://doi.org/10.1038/s41467-021-26107-z, 2021.

Ukkola, A. M. and Prentice, I. C.: A worldwide analysis of trends in water-balance evapotranspiration, Hydrol. Earth Syst. Sci., 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, J. Hydrol., 517, 1176–1187, 2014.

Van Rossum, G. and Drake, F. L.: Python 3 Reference Manual, CreateSpace, Scotts Valley, CA, ISBN 1441412697, 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 T. Knowl. Data En., 35, 8052–8072, https://doi.org/10.1109/TKDE.2022.3178128, 2022.

Wi, S. and Steinschneider, S.: Correcting the mathematical structure of a hydrological model via Bayesian data assimilation, Water Resour. Res., 58, e2022WR032123, https://doi.org/10.1029/2022WR032123, 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, J. Hydrol., 603, 127043, https://doi.org/10.1016/j.jhydrol.2021.127043, 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 Resour. Res., 57, e2020WR029471, https://doi.org/10.1029/2020WR029471, 2021.

Zeiler, M. D.: ADADELTA: An Adaptive Learning Rate Method, arXiv [preprint], https://doi.org/10.48550/arXiv.1212.5701, 2012.

Zhi, W., Ouyang, W., Shen, C., and Li, L.: Temperature outweighs light and flow as the predominant driver of dissolved oxygen in US rivers, Nature Water, 1, 249–260, 2023.

Zhong, L., Lei, H., and Gao, B.: Developing a physics-informed deep learning model to simulate runoff response to climate change in Alpine catchments, Water Resour. Res., 59, e2022WR034118, https://doi.org/10.1029/2022WR034118, 2023.