https://doi.org/10.5194/egusphere-2024-2147
Preprint. Discussion started: 30 July 2024


# Analyzing the generalization capabilities of hybrid hydrological models for extrapolation to extreme events

Eduardo Acuña Espinoza [1], Ralf Loritz [1], Frederik Kratzert [2], Daniel Klotz [3], Martin Gauch [4], Manuel Álvarez Chaves [5], Nicole Bäuerle [6], and Uwe Ehret [1]

[1]Institute of Water and Environment, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
[2]Google Research, Vienna, Austria
[3]Helmholtz Centre for Environmental Research (UFZ), Leipzig, Germany
[4]Google Research, Zurich, Switzerland
[5]Stuttgart Center for Simulation Science, Statistical Model-Data Integration, University of Stuttgart, Stuttgart, Germany
[6]Institute of Stochastics, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

**Correspondence:** Eduardo Acuña Espinoza (eduardo.espinoza@kit.edu)

**Abstract.** Data-driven techniques have shown the potential to outperform process-based models for rainfall-runoff simulation. Recently, hybrid models, which combine data-driven methods with process-based approaches, have been proposed to leverage the strengths of both methodologies, aiming to enhance simulation accuracy while maintaining certain interpretability. Expanding the set of test cases to evaluate hybrid models under different conditions, we test their generalization capabilities for extreme hydrological events, comparing their performance against Long Short-Term Memory (LSTM) networks and process-based models. Our results indicate that hybrid models show similar performance as LSTM network for most cases. However, hybrid models reported slightly lower errors in the most extreme cases, and were able to produce higher peak discharges.

## 1 Introduction

Data-driven techniques have demonstrated the potential to outperform process-based models for rainfall-runoff simulation, excelling not only in predicting average system states (Kratzert et al., 2019; Lees et al., 2021; Feng et al., 2020) but also in simulation of extreme events (Frame et al., 2022). The latter addresses concerns about the generalization capability of data-driven methods to out-of-sample conditions, showing that Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) outperformed process-based models in this type of scenario.

Recently, techniques to combine process-based models with data-driven approaches into so-called hybrid models have been proposed (Reichstein et al., 2019; Shen et al., 2023). The idea behind hybrid models is that they integrate the strengths of both process-based and data-driven approaches to improve simulation accuracy while maintaining a notion of interpretability (Jiang et al., 2020; Hoge et al., 2022). Among the various approaches available to combine methodologies, the parameterization of process-based models using data-driven techniques has shown promising results (Tsai et al., 2021). One view of this technique is that a neural network is integrated with a process-based model in an end-to-end pipeline, where the neural network handles the parameterization of the process-based model. Alternatively, this can be viewed as a neural network with a process-based head layer, which not only compresses the information into a target signal but has a certain structure that allows for the

https://doi.org/10.5194/egusphere-2024-2147
Preprint. Discussion started: 30 July 2024


recovery of untrained variables. Kraft et al. (2022) applied this method, demonstrating that substituting poorly understood or challenging-to-parameterize processes with machine learning (ML) models can effectively reduce model biases and enhance local adaptivity. Similarly, Feng et al. (2022) and Acuña Espinoza et al. (2024) employed LSTM networks to estimate the parameters of process-based models, achieving state-of-the-art performance comparable with LSTMs and outperforming stand-alone conceptual models.

In a previous study, Acuña Espinoza et al. (2024) tested the performance and interpretability of hybrid models, with the overall goal of looking at the advantages provided by adding a process-based head layer to a data-driven method. They show that hybrid models can achieve comparable performance with LSTM networks, but warn about the possibility that the data-driven section of the hybrid model compensates for structural deficiencies in the conceptual layer. Building on this research line, and expanding the set of test cases to evaluate hybrid models under different conditions, this study follows the procedure proposed by Frame et al. (2022) to investigate the ability of different models to predict out-of-sample conditions, focusing on their generalization capability to extreme events. We compare the performance of hybrid models against both traditional process-based models and stand-alone data-driven models. Our aim is to determine which model demonstrates higher predictive accuracy, particularly in simulating large hydrological events. We thereby address the following two research questions:

- How do hybrid models compare to traditional process-based and stand-alone data-driven models in the simulation of extreme hydrological events?
- Does the combination of process-based and data-driven techniques offer an advantage over stand-alone data-driven approaches?

To achieve this objective, we have structured this article as follows: Section 2 describes the training/test data split and gives an overview of the different models. In Section 3, we present the results of various tests that compare the generalization capabilities of data-driven, hybrid, and conceptual models. Lastly, Section 4 summarizes the key findings of the experiments, presents the conclusions of the study, and suggests areas for further research.

## 2 Data and methods

Donoho (2017) emphasize the importance of community benchmarks to drive model improvement. In the hydrological community, this practice has also been suggested to enable a fair comparison between new and existing methods (Shen et al., 2018; Nearing et al., 2021; Kratzert et al., 2024). Consequently, we built our experiments considering two existing studies. First, we used the procedure proposed by Frame et al. (2022) to evaluate the generalization capability of different models (see section 2.1). In accordance with this study, the experiments were conducted using the CAMELS-US dataset (Addor et al., 2017; Newman et al., 2015), in the same subset of 531 basins. Second, we used the hybrid model architecture $\delta_n(\gamma^t, \beta^t)$, further explained in section 2.2.2, proposed by Feng et al. (2022). This architecture demonstrated competitive performance with LSTM networks in their original experiments, which also used the CAMELS-US dataset.

https://doi.org/10.5194/egusphere-2024-2147
Preprint. Discussion started: 30 July 2024


### 2.1 Data handling: training/test split

To produce an out-of-sample test dataset and evaluate the generalization capability of the different models to large streamflow events, we split the training and test periods by years, based on the return period of the maximum annual discharge event. Closely following the procedure recommended by Bulletin 17C (England Jr et al., 2019), we fitted a Pearson III distribution to the annual maxima series of each basin, which we extracted from the observed CAMELS-US discharge records. We then calculated the magnitude of the discharge associated with different probabilities of exceedance. Using the discharge associated with the 5-year return period as a threshold, we classified the water years into training or test set. Figure 1 shows an example of the training/test split for basin 01054200, in the northeast of the United States. The water years (a water year is defined as the period of time between the 1st of October and the 30th of September) which contained only discharge records smaller than the associated 5-year threshold were used for training, while cases in which this threshold was exceeded were used for testing. It is important to note that there was a 365-day buffer between each training and testing period. The value of 365 days corresponds to the sequence length used by the LSTM model, and the buffer period avoids leaking test information during training. The results of the frequency analysis and the training/test data split for each basin can be found in the supplementary material accompanying this study. The original dataset contained 531 basins, each with 34 years of data (from 1980 to 2014), for a total of 18 054 years of data. After the data split process, 9 489 years were used for training, 3 429 for testing and 5 136 were buffers. Excluding the buffer data, 73% of the data was used for training and 27% for testing. This distribution is consistent with the 80%-20% theoretical split associated with the 5-year return period.

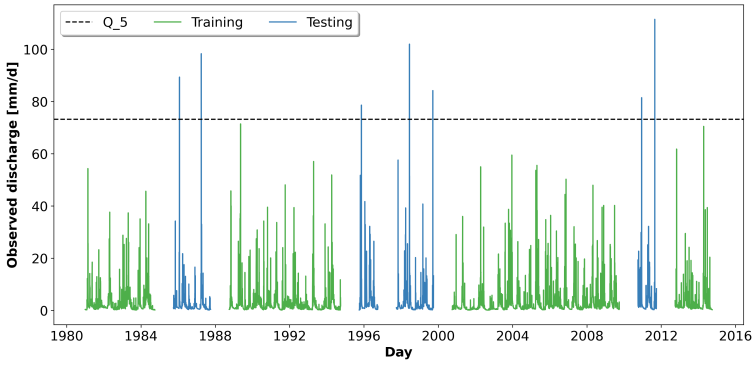


**Figure 1.** Observed discharge in mm/day from 1984 to 2016 for catchment id 01054200. Green lines represent training data, including discharge below the 5-year return period threshold, marked by the dashed line Q_5. Blue lines indicate test data for discharge exceeding this threshold. This training/test split, based on discharge exceedance probability, is designed to assess model performance under extreme hydrological conditions.

It should be noted that the results from the training/test data split differed from the ones proposed by Frame et al. (2022). In their study, the frequency analysis was done with instantaneous peak flow observations taken from the USGS NWIS, and a maximum cap of 13 water years was used to train each basin. Instead, we used the observed daily data from the CAMELS-US

https://doi.org/10.5194/egusphere-2024-2147
Preprint. Discussion started: 30 July 2024


dataset and did not impose restrictions on the maximum number of training years. We would also like to re-emphasize that this setting is meant as a form of stress-test to get an intuition of the model behavior regarding large streamflow events. In practical applications, one would not choose to use this type of setup, but one should use all available information about this kind of events for model training.

### 2.2 Data-driven, hybrid and conceptual models

The experiments in this study were conducted using three models: a stand-alone LSTM, the Hydrologiska Byråns Vattenbalansavdelning (HBV) model (Bergström, 1992) as a stand-alone conceptual model, and a hybrid approach. Both the LSTM and the hybrid model were trained using the Neural Hydrology (NH) package (Kratzert et al., 2022), while the optimization of the stand-alone conceptual model used the SPOTPY library (Houska et al., 2015). Consistent with previous studies, the LSTM and hybrid models were trained regionally, using the information from all 531 basins at the same time, while the stand-alone conceptual model was trained basin-wise (locally). In other words, in this study we compare model results of an LSTM network, a hybrid model, and 531 individually trained conceptual models.

#### 2.2.1 LSTM

The hyper-parameters for the stand-alone LSTM were taken from Frame et al. (2022). We used a single-layer LSTM with 128 hidden states, a sequence length of 365 days, a batch size of 256 and a dropout rate of 0.4. The optimization was done using the Adam algorithm (Kingma and Ba, 2014). An initial learning rate of 1e-3 was selected, which was decreased to 5e-4 and 1e-4 after 10 and 20 epochs respectively. The basin-averaged NSE loss function proposed by Kratzert et al. (2019) was used for the optimization. In a slight deviation from the original study, we trained our model for 20 epochs instead of 30. We trained our models using 5 dynamic inputs from the daymet forcing: prcp(mm/day), srad(W/m2), tmax(C), tmin(C), vp(Pa), plus the 27 static attributes listed in Table A1 of Kratzert et al. (2019).

We used an ensemble of 5 LSTM networks to produce the final simulated discharge. In other words, we trained 5 individual LSTM models, with the architecture described above, but initialized each one using a different random seed. After training, we ran each model, individually, to retrieve the simulated discharges, and took the median value as the final discharge signal that we used in the analysis. The advantage of using ensemble methods in LSTM networks was reported in Kratzert et al. (2019).

#### 2.2.2 Hybrid model: LSTM+HBV

As mentioned in the introduction, for the hybrid model architecture, we used the $\delta_n(\gamma^t, \beta^t)$ model proposed by Feng et al. (2022). In this architecture, an ensemble of 16 HBVs acting in parallel was parameterized by a single LSTM network. Each of the 16 ensemble members contained an HBV model with 4 buckets, whose flows were controlled by 11 static plus 2 time-varying parameters. The discharge of the ensemble was calculated as the mean discharge of the 16 members. Moreover, to produce the final outflow, the ensemble discharge was routed using a two-parameter unit hydrograph. In total, the LSTM

https://doi.org/10.5194/egusphere-2024-2147
Preprint. Discussion started: 30 July 2024


produced 210 parameters (13 HBV parameters*16 ensemble members + 2 routing parameters) which were used to control the ensemble of conceptual models plus the routing scheme. The model was trained end-to-end.

During training, each batch contained 256 samples, each with a sequence of 730 days. The first 365 days were used as a warmup period, to stabilize the internal states (buckets) of the HBV and reduce the effect of the initial conditions. These 365 values did not contribute to the loss function. The remaining 365 elements were used to calculate the loss, backpropagate the gradients, and update the model's weights and biases. Further details on the model implementation can be found in the *.yml* files of the supplementary material. To validate our pipeline, we benchmarked our hybrid model implementation using the experiments proposed by Feng et al. (2022). These results are shown in Appendix A, where we show a similar performance between our implementation and their results. Only after validating our pipeline, we ran the extrapolation experiments.

#### 2.2.3 Stand-alone conceptual model: HBV

As mentioned before, to have a full comparison of the model spectrum, we also included a stand-alone conceptual model. We used a single HBV model plus a unit hydrograph routing routine, resulting in a model with 14 (12 HBV + 2 routing) calibration parameters. This HBV instance has one fewer parameter (12 instead of 13) than the versions used in the hybrid model. This one parameter difference is to maintain consistency with Feng et al. (2022), where the authors used the 13-parameter HBV only when dynamic parameterization was included, and the 12-parameter model for the static version. Similar to Acuña Espinoza et al. (2024), the stand-alone conceptual models were trained basin-wise, using Shuffled Complex Evolution (SCE-UA) (Duan et al., 1994) and Differential Evolution Adaptive Metropolis (DREAM) (Vrugt, 2016), both implemented in the SPOTPY library (Houska et al., 2015). We then selected, for each basin, the calibration parameters that yielded better results.

## 3 Results and Discussion

After training the models, we performed multiple analyses to evaluate their generalization capabilities. The results of these analyses are presented in this section. All the results discussed here are for the test period.

### 3.1 Model performance comparison for whole test period

Figure 2 shows the cumulative distribution functions (CDF) for the Nash–Sutcliffe Efficiency (NSE) reported by each model, over the whole test period. We can see that the LSTM outperforms the hybrid model, reporting a median NSE of 0.75 and 0.71 respectively. Moreover, both models outperform the stand-alone HBV model, which has a median NSE value of 0.64. The implication here is that even with a different training-test split than the usual temporally contiguous subsets, our results are consistent with the ones reported by Feng et al. (2022) and Acuña Espinoza et al. (2024), where the same model ranking was achieved.

https://doi.org/10.5194/egusphere-2024-2147
Preprint. Discussion started: 30 July 2024


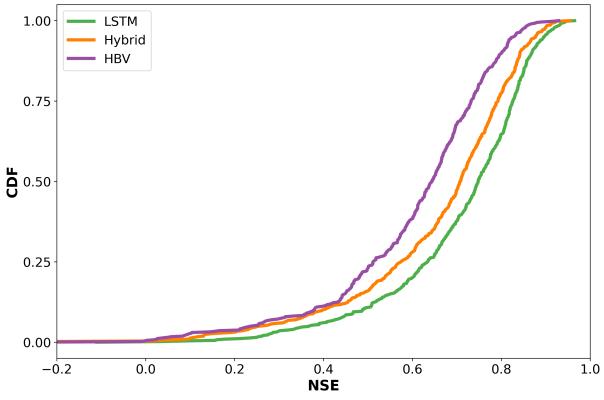


**Figure 2.** Cumulative density functions of the NSE for the different models, generated using 531 basins of the CAMELS-US dataset. The NSE was calculated over the whole test period of each basin.

### 3.2 Model performance comparison for peak flows

The metrics shown in Fig. 2 were calculated using the whole test period. Consequently, they summarized the overall performance of the three models. However, the main objective of this study is to evaluate the ability of the models to predict high-flow scenarios. To accomplish this objective, the remaining analyses were done using only peak flows.

Given the amount of data comprised in the test period (3 429 years over the 531 basins), the peak identification had to be done automatically. For this we used the *find_peaks* function of the *signal* module in the *SciPy* library (Virtanen et al., 2020), defining a 7-day window as the criterion for independent events. Moreover, we selected only the peaks above the one-year return period threshold, to have a better representation of high-flow scenarios. After we identified the peaks in the observed discharge series, we extracted the associated values from the simulated series of the different models. Figure 3a exemplifies this process for basin 01054200 in one year of the test period where each dot represents an identified peak.

Once the peaks were identified we calculated the absolute percentage error as a metric for model performance:

$$\textit{Abs percentage error} = \frac{|y_{obs} - y_{sim}|}{y_{obs}}, \tag{1}$$

where $y_{obs}$ and $y_{sim}$ are the observed and simulated discharge, respectively. Figure 3b presents, for each model, the distribution of the absolute percentage error for all the peak flows. This figure shows a similar distribution for the three models, with the LSTM presenting a slightly lower median error than the hybrid and stand-alone HBV. The finding that LSTMs outperformed process-based models aligns with Frame et al. (2022) and helps to challenge the notion that data-driven methods are less capable of extrapolation (Reichstein et al., 2019; Slater et al., 2023). In the case of the hybrid model, although the LSTM exhibits a slightly lower median error, the error distributions of both models are similar. This trend will also be observed in other experiments discussed in the following sections. Therefore, we do not find strong evidence suggesting that one architecture is significantly better than the other in this scenario, leaving it to the reader's discretion to choose the model that best suits their needs.

https://doi.org/10.5194/egusphere-2024-2147
Preprint. Discussion started: 30 July 2024


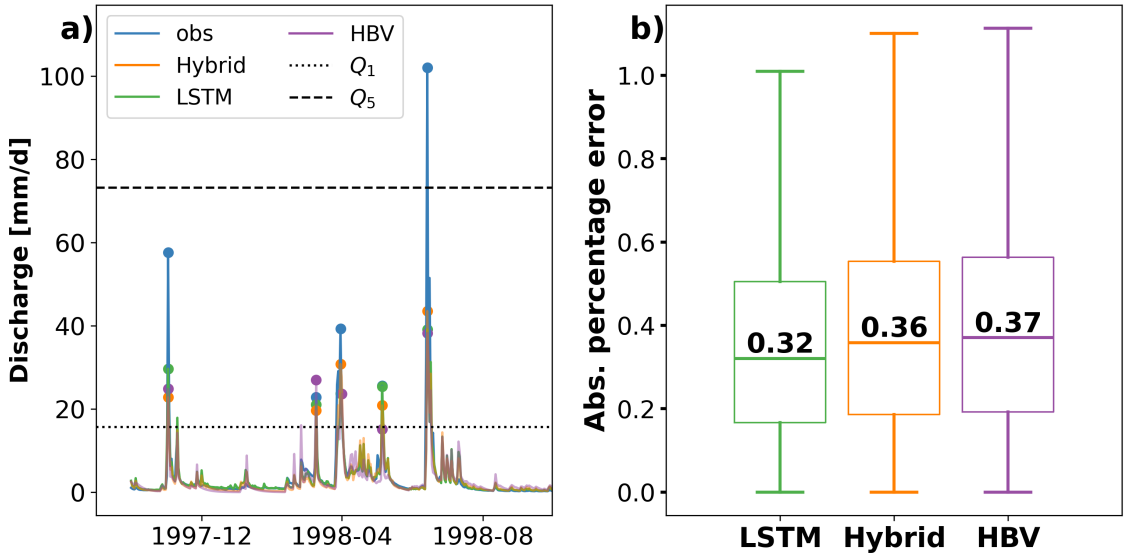


**Figure 3.** a) Example of peak identification for basin 01054200 in one year of the test period. $Q_1$ represent the one-year return period threshold, which was used to identify peak events. $Q_5$ represent the five-year return period threshold, which was used for the training/testing data split b) Absolute percentage error between the observed peak discharge and the associated simulation value for the different models. The results of subplot b present the error distribution, from all 531 basins, calculated only for the peak flows of their test period (total of 17580 values).

### 3.3 Model performance comparison for out-of-sample peak flows

Figure 3b allowed us to evaluate the performance of the models only in peak discharges. However, this still did not give us a performance metric for values exclusively outside of the training range. More specifically, the results presented in Fig. 3b evaluated the error in 17 580 observed events. Considering that the test period contained 3 429 years, we got an average of 5 peaks per year. However, as shown in Fig. 3a, these peaks were not necessarily larger than the 5-year return period thresholds used during training. Figure 4 shows the same error metric but classifies the peaks based on their return period. The four categories to the right of the dashed vertical line present the errors associated with discharge outside of the training range, giving a strict evaluation of the generalization capabilities of the models. We can see that the LSTM slightly outperformed the hybrid and HBV models in the 1-5 and 5-25 return period. In the remaining three intervals the performance of the LSTM and hybrid are comparable, with the HBV also showing similar behavior for the last two. As shown in Appendix B, the differences in the median values of the last three categories are within the range of statistical noise. Figure B1 shows that if we use the individual predictions of the 5 members of the LSTM ensemble, instead of reporting the median signal, the ranking of these three categories can change. Consequently, we can conclude that for higher return periods all models perform similarly.

In most cases, the errors increased for higher return periods. This was expected as models were trying to generalize to flows farther away from their training range. On the other hand, the 100+ return period peaks presented similar or slightly lower

https://doi.org/10.5194/egusphere-2024-2147
Preprint. Discussion started: 30 July 2024


errors than the ones in the 50-100 category. At this point, the reported errors were close to 60%, which indicated that no model could satisfactorily reproduce the observed peaks. Moreover, because of the characteristic of the metric (see equation 1), the error was scaled by the magnitude of the observation, which would explain why the 50-100 and 100+ presented similar errors.

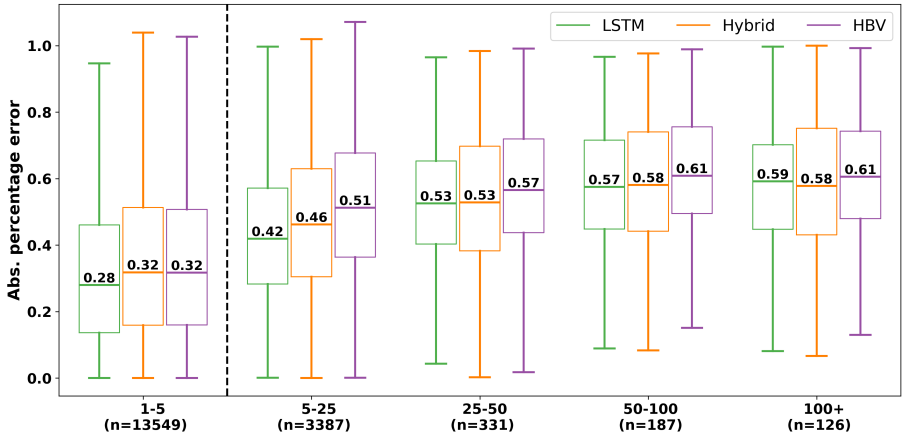


**Figure 4.** Absolute percentage error between the observed peak discharge and the associated simulation value for the different models, classified by the return period of the observed peaks. The four categories to the right of the dashed vertical line present the errors associated with observed discharge above the 5-year return period threshold, evaluating the out-of-sample capabilities of the models. The n-value below each category indicates the amount of data used to produce the box-plot.

### 3.4 Saturation analysis: behavior of the models during extreme flow scenarios

Kratzert et al. (2024) explain that due to the LSTM model architecture, there is a theoretical prediction limit for this type of model, which is a function of the weights and bias of the head linear layer, and consequently is defined during the training process. In other words, independently of the input series, the associated prediction cannot go above the theoretical limit. On the other hand, both the hybrid and the stand-alone HBV model do not have such a theoretical limit. The conceptual model architecture is defined with an unlimited capacity in the buckets, and due to mass conservation, all the water received by the models, after evapotranspiration and other abstractions, has to go out. Therefore, as a further analysis, we evaluated the behaviour of the model in the overall highest events.

First, we selected the 531 highest peaks in the test period. These peaks were selected as the overall highest events and came from 171 out of the 531 basins. Therefore, not all basins participated in this analysis and the number of peaks that came out of each basin did not have to be the same. Then, we selected the respective simulated values and plotted their CDF in Fig. 5a.

From this figure, we can see that all three models underestimate the peak discharges. This phenomenon as such is not necessarily an indication of model deficiencies. All models provide an estimation of the expected streamflow for a given timestep. However, since we selected the highest discharges across all basins for this exercise, it stands to reason that we selected observations from the upper-quantiles of the respective conditional distributions. Figure 5a shows that the LSTM

https://doi.org/10.5194/egusphere-2024-2147
Preprint. Discussion started: 30 July 2024


produces the lower maximum among the three models, and that the HBV and hybrid models present a similar CDF. The similarity between the last two indicates that in extreme regimes, the data-driven part of the hybrid model does not contribute much, and the conceptual part is the one that allows the model to produce higher flows than the stand-alone LSTM. Moreover, we can also see that all three models do not produce values larger than the maximum value they saw during training (dashed vertical line in Fig. 5a). Figure 5b shows the error distribution for the same 531 highest peaks. The three models present a similar median error, with the hybrid and HBV models reporting slightly longer tails towards smaller absolute percentage errors.

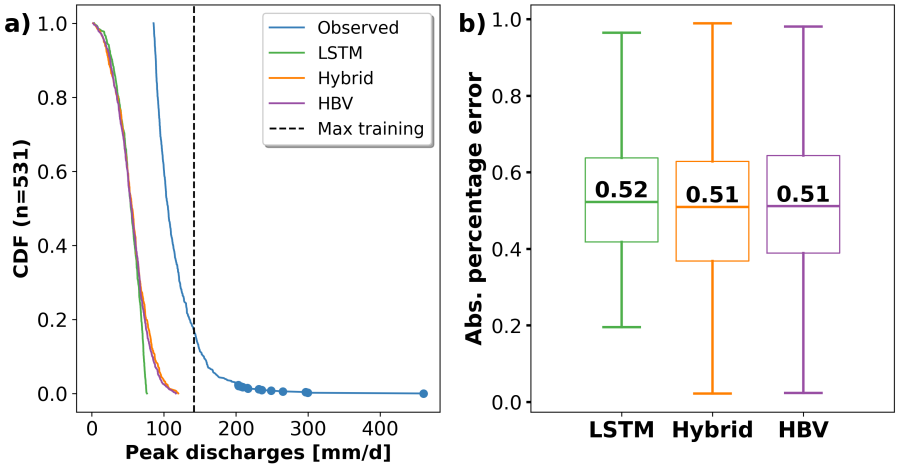


**Figure 5.** a) CDF of the 531 observed highest discharge values across all basins and their respective simulated values. The blue dots help visualize that under 3% of the events have values between 200 and 400 mm/day. b) Absolute percentage error of the 531 highest discharges for the different models.

The saturation problem in LSTM models with a single linear head layer, as described by Kratzert et al. (2024), arises due to the inherent limitations of the model architecture, resulting in a theoretical prediction limit (see equation B2 of Kratzert et al. (2024)). From our experiments, we can see that the maximum value predicted by an LSTM instance was 78.9 mm/day which is close to the theoretical prediction limit, which we calculated to be 83.9 mm/day. The other four instances of the ensemble reported similar results. The theoretical prediction limit is a function of the weights and bias of the head layer, which are a result of the training process. In our experiment we artificially restricted the training data to discharge smaller than the 5-year return period thresholds, reducing the support of the data space the model was fitted to. Consequently, this setup directly intensified the saturation problem. In practical applications where the model is trained on all available data, the saturation issue would tend to decrease in relevance. However, a theoretical saturation limit remains, which is an undesirable property in a hydrological model, especially in cases where we are designing infrastructure for extreme events outside of any training data (e.g., 1 000 year flood). Further research should be invested in overcoming this problem.

https://doi.org/10.5194/egusphere-2024-2147
Preprint. Discussion started: 30 July 2024


Apart from the statistical artifacts introduced by our selection procedure, we found two potential issues that might lead to the peak underestimation of the hybrid models. Figure 6 shows the precipitation and observed discharge, together with the accumulated value and the simulated discharge series, for 4 of the largest events in the dataset.

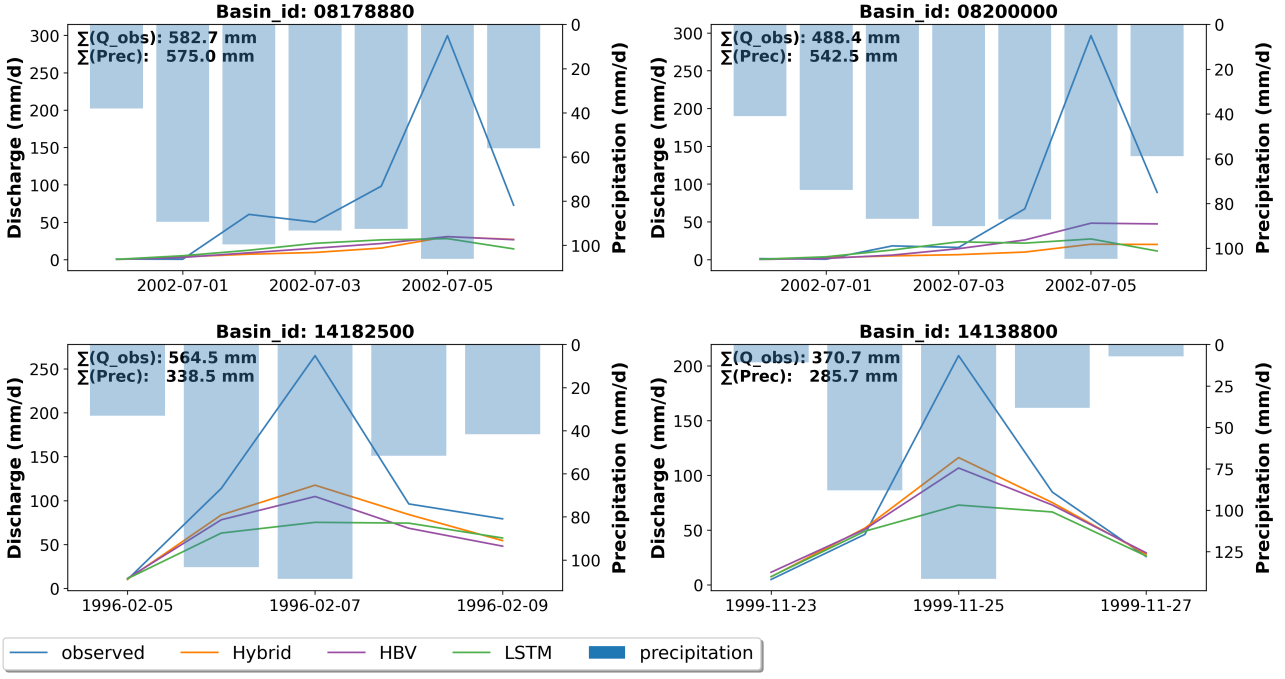


**Figure 6.** Example of 4 of the most extreme events presented in the dataset. The subplots show the precipitation series and the observed and simulated hydrographs. ∑Q_obs and ∑Prec indicate the cumulative sum of the discharge and precipitation series. Basins 08178880 and 08200000 have similar precipitation and discharge volumes. Basins 14182500 and 14138800 have a precipitation volume smaller than the discharge volume. The dates of the events are in format YYYY-MM-DD.

For basins 08178880 and 08200000, the accumulated precipitation of the event is similar to or larger than the accumulated discharge, however, the simulated series strongly underestimates the discharge. This behaviour can arise due to structural limitations in the hydrological model. For example, given the lack of a fast response channel, a high precipitation pulse can be divided and routed through several linear reservoirs, attenuating the respective discharge peak. This effect could have been strengthened by our training/test split, because the optimization parameters, which control the interaction between the buckets, were learned for certain conditions, which were inadequate for other out-of-sample hydrological events.

In this regard, the hybrid model presents a theoretical advantage over HBV through the possibility of dynamic parameterization that adapts the model behaviour to current conditions. The $\delta_n(\gamma^t, \beta^t)$ hybrid model uses a dynamic $\beta$ coefficient to control the recharge rate at which precipitation was transferred to the other buckets. We noticed that during high-intensity events the $\beta$ value reached the limits of their predefined interval, which limited the model to further adapt its behaviour.

https://doi.org/10.5194/egusphere-2024-2147
Preprint. Discussion started: 30 July 2024


The second issue that we found is a possible bias in the input data. For basins 14182500 and 14138800, the accumulated precipitation is smaller than the accumulated discharge, which would explain the underestimation of the simulated values. Given that these two basins are located in the state of Oregon (north-west of the USA), and accounting for the dates of both events, there is also the possibility that the high discharge is partially caused by snowmelt. In this case, the precipitation mass would need to be corrected. Nevertheless, the snow module of HBV is not reproducing this behaviour, which again points towards a structural deficiency in the model.

### 3.5 Limitations and uncertainties

The comparison that we presented here was done using the hybrid architecture proposed by Feng et al. (2022). This architecture was chosen because it gave a competitive performance with LSTM in their original experiment and because the code was open source. Other hybrid model architectures might give different results, and we encourage the hydrological community to expand the test cases presented here.

Moreover, and as we indicated before, the training/test split was intended as a form of stress-testing, to get an intuition of the model when generalizing to unseen events. For the reasons stated in previous sections, this stress-testing method directly affects the saturation problem in the LSTM and the parameter optimization for the hybrid and conceptual models. In a practical case, one should use all the data during model training, to increase the performance of the models.

Lastly, differences between simulated and observed values, especially in extreme events, can also be attributed to higher uncertainty in the observed quantities, including discharge and precipitation (Di Baldassarre and Montanari, 2009; Westerberg and McMillan, 2015; Bárdossy and Anwar, 2023). We did not consider this type of uncertainty in our analysis, as this would be outside of the scope of the paper.

## 4 Summary and conclusions

In this study, we evaluated the generalization capabilities of data-driven, hybrid, and conceptual models for predicting extreme hydrological events. Following the methodology proposed by Frame et al. (2022), we partitioned our data based on the occurrence frequencies, using the 5-year return period discharge as a threshold. We trained our models using information from water years with discharges strictly lower than the threshold and tested their performance on low-probability data. This setting was meant as a form of stress-test to get an intuition of the model behavior regarding large streamflow events. Our findings indicated that the LSTM slightly outperforms the hybrid and HBV models for 1-5 and 5-25 return periods, and all models show similar performance for higher discharges.

While all models underestimated extreme flow scenarios, the hybrid model and HBV model were able to simulate higher discharges than the LSTM model. Upon further investigation, we noticed that the reasons for underestimating the extreme flow scenarios were different. The LSTM had a theoretical limit due to its architecture, and even though in practice this problem can be attenuated it is not a desirable property for a hydrological model. Additional research to overcome this limitation should be encouraged. On the other hand, the hybrid and HBV models underestimated the discharge due to structural deficiencies and

https://doi.org/10.5194/egusphere-2024-2147
Preprint. Discussion started: 30 July 2024


possible bias in the input data. The dynamic parameterization of hybrid models might help reduce the former, by changing the model response based on current conditions. This idea is conceptually similar to how an LSTM operates, in which the gate structures operate based on current and past conditions.

Overall, for the experiments performed here, we did not find strong evidence suggesting that there is a significant difference between the extrapolation capabilities of LSTM networks and hybrid models, and we leave it to the reader's discretion to choose the model that best suits their needs.

*Code availability.* The code used to conduct all the analyses in this paper are publicly available at https://github.com/eduardoAcunaEspinoza/hybrid_extrapolation/tree/v0.1

*Data availability.* The CAMELS US dataset is freely available at https://doi.org/10.5065/D6MW2F4D (Newman et al., 2022). All the data generated for this publication can be found at https://doi.org/10.5281/zenodo.12705219

## Appendix A: Benchmarking hybrid model

As explained in the manuscript, for our hybrid model we used the $\delta_n(\gamma^t, \beta^t)$ architecture proposed by Feng et al. (2022). Because our experiment pipeline was executed in the NeuralHydrology package, we first had to benchmark our model implementation against the original case. Figure A1 shows that our model implementation produced similar results to the one reported by Feng.

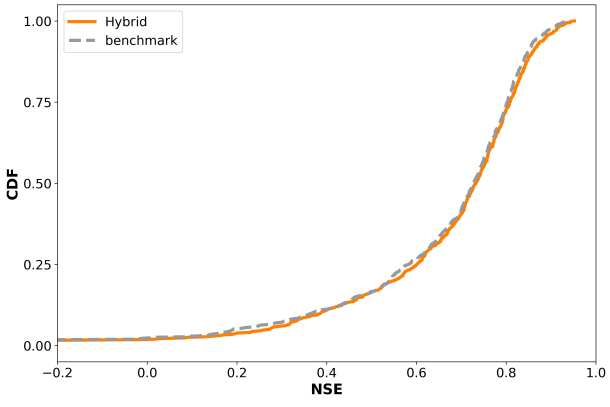


**Figure A1.** Cumulative density function of the NSE for different models, generated using 671 basins of the CAMELS-US dataset.

https://doi.org/10.5194/egusphere-2024-2147
Preprint. Discussion started: 30 July 2024


## Appendix B: Effect on absolute percentage error due to random initialization of the LSTM model

Figure B1 shows the effect of different model initializations, for the LSTM network, in the absolute percentage error metric. The ranking of the models in the last three categories (25-50, 50-100, 100+) varies depending on the model initialization. This indicates that the differences in the median values are within the statistical noise, and we cannot conclude that one model is better than the other.

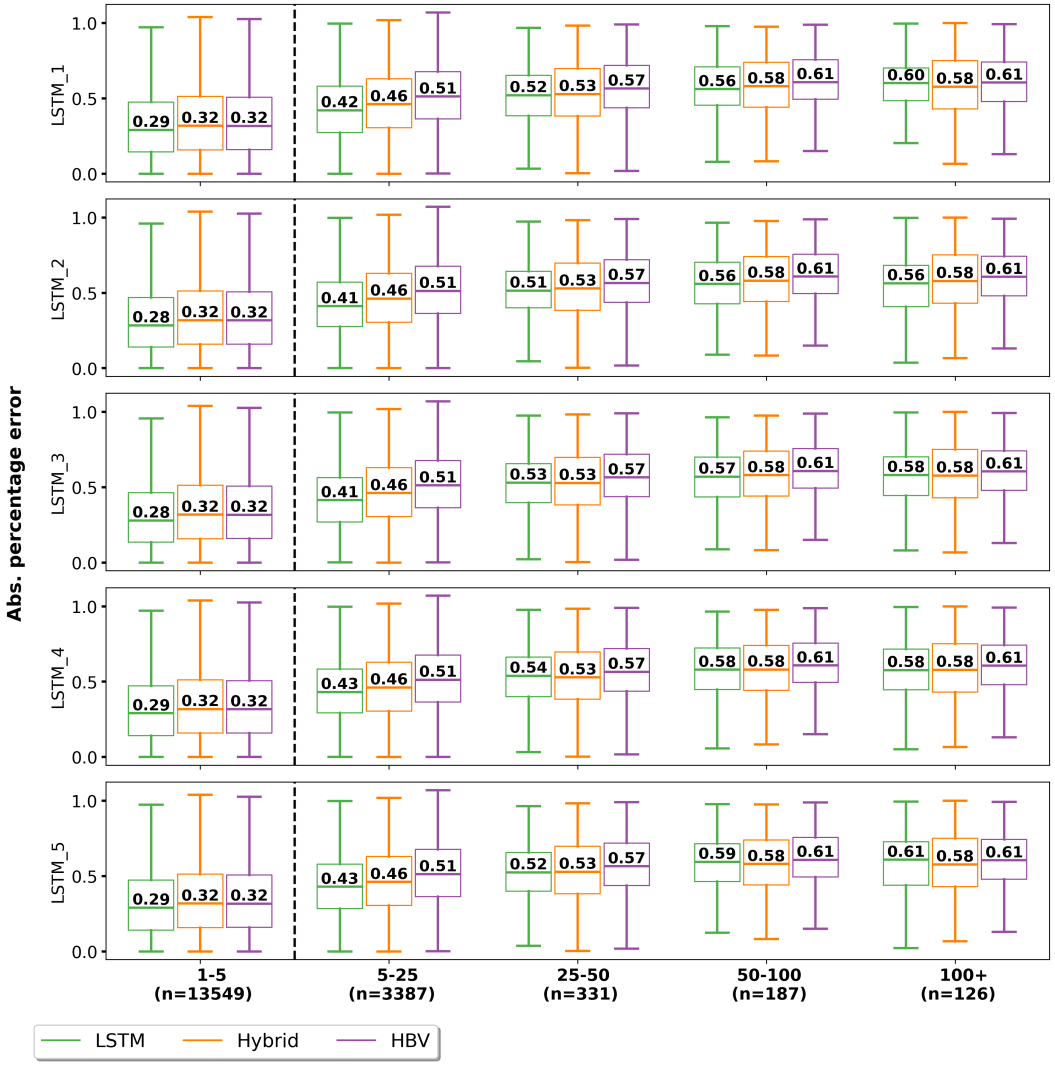


**Figure B1.** Variation in absolute percentage error due to random initialization of the LSTM model.

https://doi.org/10.5194/egusphere-2024-2147
Preprint. Discussion started: 30 July 2024


*Author contributions.* The original idea of the manuscript was developed by all authors. The codes were written by E.A.E with support from F.K., M.G., and D.K. The simulations were conducted by E.A.E. Results were further discussed by all authors. The draft of the manuscript was prepared by E.A.E. Reviewing and editing was provided by all authors. Funding was acquired by U.E. and N.B. All authors have read and agreed to the current version of the manuscript.

*Competing interests.* Some authors are members of the editorial board of HESS.

*Acknowledgements.* We would like to thank the Google Cloud Program (GCP) team, for awarding us credits to support our research and run the models.

*Financial support.* This project has received funding from the KIT Center for Mathematics in Sciences, Engineering and Economics under the seed funding program.

https://doi.org/10.5194/egusphere-2024-2147
Preprint. Discussion started: 30 July 2024


## References

Acuña Espinoza, E., Loritz, R., Álvarez Chaves, M., Bäuerle, N., and Ehret, U.: To Bucket or not to Bucket? Analyzing the performance and interpretability of hybrid hydrological models with dynamic parameterization, Hydrology and Earth System Sciences, 28, 2705–2719, https://doi.org/10.5194/hess-28-2705-2024, 2024.

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

Bárdossy, A. and Anwar, F.: Why do our rainfall–runoff models keep underestimating the peak flows?, Hydrology and Earth System Sciences, 27, 1987–2000, https://doi.org/10.5194/hess-27-1987-2023, 2023.

Bergström, S.: THE HBV MODEL - its structure and applications, Tech. rep., Sveriges Meteorologiska Och Hydrologiska Institut, https://www.smhi.se/en/publications/the-hbv-model-its-structure-and-applications-1.83591, 1992.

Di Baldassarre, G. and Montanari, A.: Uncertainty in river discharge observations: a quantitative analysis, Hydrology and Earth System Sciences, 13, 913–921, https://doi.org/10.5194/hess-13-913-2009, 2009.

Donoho, D.: 50 years of data science, Journal of Computational and Graphical Statistics, 26, 745–766, https://doi.org/10.1080/10618600.2017.1384734, 2017.

Duan, Q., Sorooshian, S., and Gupta, V. K.: Optimal use of the SCE-UA global optimization method for calibrating watershed models, Journal of Hydrology, 158, 265–284, https://doi.org/https://doi.org/10.1016/0022-1694(94)90057-4, 1994.

England Jr, J. F., Cohn, T. A., Faber, B. A., Stedinger, J. R., Thomas Jr, W. O., Veilleux, A. G., Kiang, J. E., and Mason Jr, R. R.: Guidelines for determining flood flow frequency—Bulletin 17C, Tech. rep., US Geological Survey, https://doi.org/10.3133/tm4B5, 2019.

Feng, D., Fang, K., and Shen, C.: Enhancing Streamflow Forecast and Extracting Insights Using Long-Short Term Memory Networks With Data Integration at Continental Scales, Water Resources Research, 56, e2019WR026793, https://doi.org/https://doi.org/10.1029/2019WR026793, 2020.

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 Resources Research, 58, e2022WR032404, https://doi.org/https://doi.org/10.1029/2022WR032404, 2022.

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

Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, Neural Computation, 9, 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735, 1997.

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

Houska, T., Kraft, P., Chamorro-Chavez, A., and Breuer, L.: SPOTting model parameters using a ready-made python package, PloS one, 10, e0145180, https://doi.org/10.1371/journal.pone.0145180, 2015.

Jiang, S., Zheng, Y., and Solomatine, D.: Improving AI system awareness of geoscience knowledge: Symbiotic integration of physical approaches and deep learning, Geophysical Research Letters, 47, e2020GL088229, https://doi.org/10.1029/2020GL088229, 2020.

Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980, https://doi.org/10.48550/arXiv.1412.6980, 2014.

https://doi.org/10.5194/egusphere-2024-2147
Preprint. Discussion started: 30 July 2024


Kraft, B., Jung, M., Körner, M., Koirala, S., and Reichstein, M.: Towards hybrid modeling of the global hydrological cycle, Hydrology and Earth System Sciences, 26, 1579–1614, https://doi.org/10.5194/hess-26-1579-2022, 2022.

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, Hydrology and Earth System Sciences, 23, 5089–5110, https://doi.org/10.5194/hess-23-5089-2019, 2019.

Kratzert, F., Gauch, M., Nearing, G., and Klotz, D.: NeuralHydrology — A Python library for Deep Learning research in hydrology, Journal of Open Source Software, 7, 4050, https://doi.org/10.21105/joss.04050, 2022.

Kratzert, F., Gauch, M., Klotz, D., and Nearing, G.: HESS Opinions: Never train an LSTM on a single basin, Hydrology and Earth System Sciences Discussions, 2024, 1–19, https://doi.org/10.5194/hess-2023-275, 2024.

Lees, T., Buechel, M., Anderson, B., Slater, L., Reece, S., Coxon, G., and Dadson, S. J.: Benchmarking data-driven rainfall-runoff models in Great Britain: a comparison of long short-term memory (LSTM)-based models with four lumped conceptual models, Hydrology and Earth System Sciences, 25, 5517–5534, https://doi.org/10.5194/hess-25-5517-2021, 2021.

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., and Gupta, H. V.: What role does hydrological science play in the age of machine learning?, Water Resources Research, 57, e2020WR028 091, https://doi.org/10.1029/2020WR028091, 2021.

Newman, A., Sampson, K., Clark, M., Bock, A., Viger, R. J., Blodgett, D., Addor, N., and Mizukami, M.: CAMELS: Catchment Attributes and MEteorology for Large-sample Studies. Version 1.2, https://gdex.ucar.edu/dataset/camels.html, accessed 10 Jul 2024, 2022.

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

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., and Carvalhais, N.: Deep learning and process understanding for data-driven Earth system science, Nature, 566, 195–204, https://doi.org/10.1038/s41586-019-0912-1, 2019.

Shen, C., Laloy, E., Elshorbagy, A., Albert, A., Bales, J., Chang, F., Ganguly, S.and Hsu, K., Kifer, D., Fang, Z., Dongfeng, L., and Xiaodong, Li.and Tsai, W.: HESS Opinions: Incubating deep-learning-powered hydrologic science advances as a community, Hydrology and Earth System Sciences, 22, 5639–5656, https://doi.org/10.5194/hess-22-5639-2018, 2018.

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., Rahmani, F., Song, Y., Beck, H. E., Bindas, T., Dwivedi, D., Fang, K., Höge, M., Rackauckas, C., Mohanty, B., Roy, T., Xu, C., and Lawson, K.: Differentiable modelling to unify machine learning and physical models for geosciences, Nature Reviews Earth & Environment, 4, 552–567, https://doi.org/10.1038/s43017-023-00450-9, 2023.

Slater, L. J., Arnal, L., Boucher, M.-A., Chang, A. Y.-Y., Moulds, S., Murphy, C., Nearing, G., Shalev, G., Shen, C., Speight, L., Villarini, G., Wilby, R. L., Wood, A., and Zappa, M.: Hybrid forecasting: blending climate predictions with AI models, Hydrology and Earth System Sciences, 27, 1865–1889, https://doi.org/10.5194/hess-27-1865-2023, 2023.

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, Nature Communications, 12, 5988, https://doi.org/10.1038/s41467-021-26107-z, 2021.

https://doi.org/10.5194/egusphere-2024-2147
Preprint. Discussion started: 30 July 2024


Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, I., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., SciPy 1.0 Contributors, Vijaykumar, A., Bardelli, A. P., Rothberg, A., Hilboll, A., Kloeckner, A., Scopatz, A., Lee, A., Rokem, A., Woods, C. N., Fulton, C., Masson, C., Häggström, C., Fitzgerald, C., Nicholson, D. A., Hagen, D. R., Pasechnik, D. V., Olivetti, E., Martin, E., Wieser, E., Silva, F., Lenders, F., Wilhelm, F., Young, G., Price, G. A., Ingold, G.-L., Allen, G. E., Lee, G. R., Audren, H., Probst, I., Dietrich, J. P., Silterra, J., Webber, J. T., Slavič, J., Nothman, J., Buchner, J., Kulick, J., Schönberger, J. L., de Miranda Cardoso, J. V., Reimer, J., Harrington, J., Rodríguez, J. L. C., Nunez-Iglesias, J., Kuczynski, J., Tritz, K., Thoma, M., Newville, M., Kümmerer, M., Bolingbroke, M., Tartre, M., Pak, M., Smith, N. J., Nowaczyk, N., Shebanov, N., Pavlyk, O., Brodtkorb, P. A., Lee, P., McGibbon, R. T., Feldbauer, R., Lewis, S., Tygier, S., Sievert, S., Vigna, S., Peterson, S., More, S., Pudlik, T., Oshima, T., Pingel, T. J., Robitaille, T. P., Spura, T., Jones, T. R., Cera, T., Leslie, T., Zito, T., Krauss, T., Upadhyay, U., Halchenko, Y. O., and Vázquez-Baeza, Y.: SciPy 1.0: fundamental algorithms for scientific computing in Python, Nature Methods, 17, 261–272, https://doi.org/10.1038/s41592-019-0686-2, 2020.

Vrugt, J. A.: Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation, Environmental Modelling and Software, 75, 273–316, https://doi.org/https://doi.org/10.1016/j.envsoft.2015.08.013, 2016.

Westerberg, I. K. and McMillan, H. K.: Uncertainty in hydrological signatures, Hydrology and Earth System Sciences, 19, 3951–3968, https://doi.org/10.5194/hess-19-3951-2015, 2015.