Hydrol. Earth Syst. Sci., 29, 1939–1962, 2025
https://doi.org/10.5194/hess-29-1939-2025

Hydrology and Earth System Sciences
Open Access
EGU

Long short-term memory networks for enhancing real-time flood forecasts: a case study for an underperforming hydrologic model

Sebastian Gegenleithner[1,2,★], **Manuel Pirker**[1,★], **Clemens Dorfmann**[2], **Roman Kern**[3], **and Josef Schneider**[1]

[1]Institute of Hydraulic Engineering and Water Resources Management, Graz University of Technology, Stremayrgasse 10/II, 8010 Graz, Austria
[2]flow engineering, Lessingstraße 30, 8010 Graz, Austria
[3]Institute of Interactive Systems and Data Science, Graz University of Technology, Sandgasse 36/III, 8010 Graz, Austria
★These authors contributed equally to this work.

Correspondence: Sebastian Gegenleithner (s.gegenleithner@gmail.com) and Manuel Pirker (manuel.pirker@tugraz.at)

Received: 8 April 2024 – Discussion started: 14 May 2024
Revised: 2 November 2024 – Accepted: 17 January 2025 – Published: 17 April 2025

Abstract. Flood forecasting systems play a key role in mitigating socioeconomic damage caused by flood events. The majority of these systems rely on process-based hydrologic models (PBHMs), which are used to predict future runoff. Many operational flood forecasting systems additionally implement models aimed at enhancing the predictions of the PBHM, either by updating the PBHM's state variables in real time or by enhancing its forecasts in a post-processing step. For the latter, autoregressive integrated moving average (ARIMA) models are frequently employed. Despite their high popularity in flood forecasting, studies have pointed out potential shortcomings of ARIMA-type models, such as a decline in forecast accuracy with increasing lead time. In this study, we investigate the potential of long short-term memory (LSTM) networks for enhancing the forecast accuracy of an underperforming PBHM and evaluate whether they are able to overcome some of the challenges presented by ARIMA models. To achieve this, we developed two hindcast–forecast LSTM models and compared their forecast accuracies to that of a more conventional ARIMA model. To ensure comparability, one LSTM was restricted to use the same data as ARIMA (eLSTM), namely observed and simulated discharge, while the other additionally incorporated meteorologic forcings (PBHM-HLSTM). Considering the PBHM's poor performance, we further evaluated if the PBHM-HLSTM was able to extract valuable information from the PBHM's results by analyzing the relative importance of each input feature. Contrary to ARIMA, the LSTM networks were able to mostly sustain a high forecast accuracy for longer lead times. Furthermore, the PBHM-HLSTM also achieved a high prediction accuracy for flood events, which was not the case for ARIMA or the eLSTM. Our results also revealed that the PBHM-HLSTM relied, to some degree, on the PBHM's results, despite its mostly poor performance. Our results suggest that LSTM models, especially when provided with meteorologic forcings, offer a promising alternative to frequently employed ARIMA models in operational flood forecasting systems.

1 Introduction

Floods are among the most common and most destructive natural disasters around the world (Yaghmaei et al., 2020). Alongside other mitigation measures, flood forecasting systems play a key role in increasing resilience to such events. In principle, flood forecasting systems enable the prediction of future river discharge, thereby empowering decision-makers and emergency forces with respect to the implementation of effective early countermeasures in the case of flooding events. Examples of such flood forecasting systems are given by Werner et al. (2009), Addor et al. (2011), Nester et al. (2016), Borsch et al. (2021), or Nearing et al. (2024).

To date, most operational flood forecasting systems are built around process-based hydrologic models (PBHMs). These models predict future river discharge by utilizing conceptual or more physically based approaches that depict the

individual components of the hydrologic cycle in the catchment. In recent years, many researchers have proposed solely data-driven models as an alternative to PBHMs. Particularly, models based on long short-term memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) have gained recognition for their capability to accurately model river discharge. For example, Kratzert et al. (2019b) demonstrated that their LSTM model was able to outperform two PBHMs across not only multiple gauged catchments but also in ungauged catchments. Although data-driven models have proven to be a viable alternative to PBHMs for modeling river discharge, they are still rarely applied as the core component in operational flood forecasting systems (Nevo et al., 2022).

The primary task of PBHMs employed in operational flood forecasting systems is predicting a sequence of future discharge values. The length of this sequence is chosen based on the characteristics of the catchment and is referred to as the forecast horizon. For the chosen forecast horizon, the PBHM derives the discharge forecasts based on meteorologic quantities as well as its current system state at the beginning of the forecast horizon, e.g., the state of the snow cover, the soil moisture, or the available water below and above the surface (river discharge). A common practice in flood forecasting is to use real-time observations of these state variables, evaluate how the model was able to replicate them in the past, and use this knowledge to enhance the model's forecasts. Considering the available literature, the most relevant forecast-enhancing strategies can be grouped as follows:

1. *State updating*. The basic idea behind this concept is to use observational data to update parts of the hydrologic model in real time, allowing it to more accurately reflect the true state of the system. Commonly applied methods for state updating in flood forecasting include variants of the Kalman filter (Kalman, 1960) or particle filters (as demonstrated by Weerts and El Serafy, 2006).

2. *Error correction*. These methods use observations of one or multiple state variables, mostly river discharge, to enhance the hydrologic model's forecasts in a post-processing step. Specifically, models belonging to the autoregressive integrated moving average (ARIMA) family are frequently employed for this purpose. However, despite their high popularity, numerous studies have pointed out the potential limitations of these models in hydrologic modeling applications.

Firstly, ARIMA models often exhibit a decline in forecast accuracy with increasing lead time. For instance, Brath et al. (2002) demonstrated that the forecast accuracy of an adaptively updated ARIMA-type model degraded to match the accuracy of the not-updated model after six time steps. A less significant performance decrease was observed for an ARIMA-type model that was calibrated with a split-sample strategy. Similarly, Broersen and Weerts (2005) demonstrated that their employed ARIMA-type models were able to significantly increase the prediction accuracy within the first day; however, for more temporally distant predictions, only slight differences were found to forecasts corrected with the mean discharge over the last 3 weeks. Secondly, ARIMA models struggle to provide accurate forecasts for flood events when the underlying hydrologic model fails to give an adequate initial estimation, as shown by studies such as Liu et al. (2015). In their study, Liu et al. (2015) assessed the predictive skill of an ARIMA-corrected PBHM for a total of four significant flood events. While their model demonstrated a high forecast accuracy for events that were already captured well by the hydrologic model, it failed in one instance where this was not the case. Reasonable forecasts for this event could only be obtained in the consecutive forecast step, followed by a rapid decline in forecast accuracy.

Recently, researchers have explored the potential of neural networks, particularly recurrent neural networks (RNNs), to enhance the results obtained from PBHMs, and the outcomes have been remarkably successful. Although the focus of their study was on model diagnostics, Rozos et al. (2021) demonstrated that RNNs and LSTM networks, trained on meteorologic data and the PBHM's output, have the potential to enhance the model accuracy of underperforming PBHMs. In a large-sample study, Konapala et al. (2020) tested various LSTM variants to enhance the prediction accuracy of a PBHM. They found that, overall, their hybrid LSTM models that incorporated the results of the PBHM outperformed both the PBHM and, in most instances, a standalone LSTM. They also found that the highest improvements were achieved for catchments where the PBHM was underperforming. A comparable study was also conducted by Frame et al. (2021). In their study, the authors showed that discharge predictions could be improved by LSTM models that incorporated the results of the PBHM. However, they also demonstrated that these models, in many instances, were outperformed by a standalone LSTM that did not incorporate information obtained by the PBHM.

Given the promising findings of the aforementioned studies, we recognize the substantial potential of neural networks to enhance the forecast accuracy of underperforming PBHMs employed in operational flood forecasting systems. Especially in aspects where ARIMA correction methods have previously demonstrated shortcomings, such as maintaining a high forecast accuracy for longer lead times or accurately correcting poor flood event predictions, neural networks might yield more accurate forecasts. To test this hypothesis, we developed two LSTM model variants, both implemented with a hindcast–forecast architecture, similar to that presented by Gauch et al. (2021) or Nevo et al. (2022), and compared their forecast performance to that of a conventional ARIMA model. To ensure comparability, one LSTM variant, eLSTM, was restricted to use the same input data as ARIMA, specifically observed discharge and that obtained by the PBHM. The second variant, PBHM-HLSTM, was implemented with the same architecture as the eLSTM but addi-

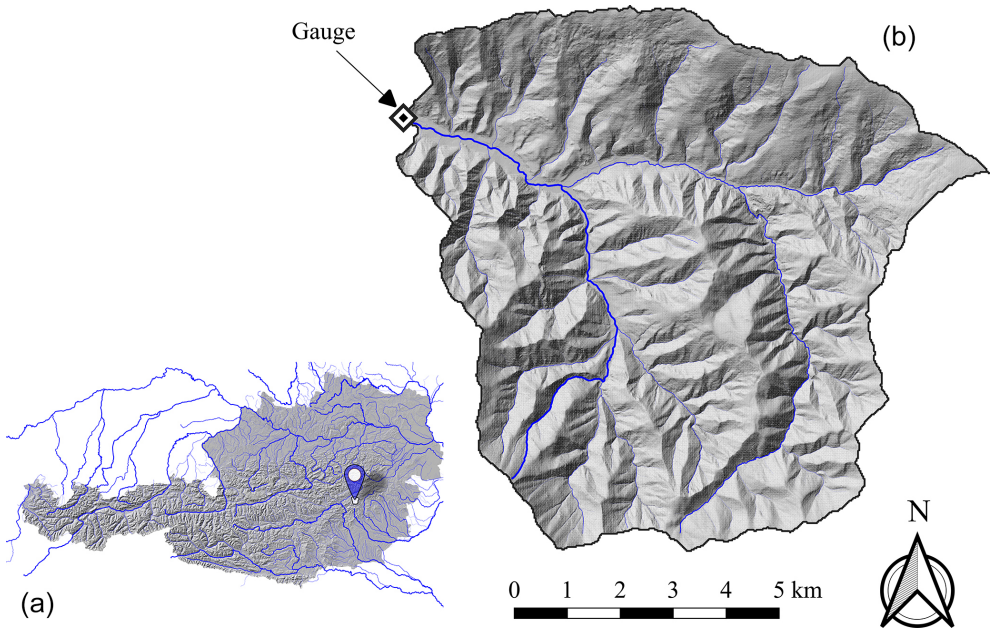

Figure 1. (a) Location of the study catchment in Austria. **(b)** Outline of the study catchment (black line) including the gauging station (black-and-white diamond) and the main river network (blue lines). This figure was created using the following datasets: Umweltbundesamt GmbH (2022) and Land Kärnten (2019).

tionally incorporated meteorologic forcings. It has to be mentioned that our ARIMA model relied on forecasting residuals, whereas both LSTM variants directly predicted future discharge. For the PBHM-HLSTM exclusively, we also evaluated the contribution, i.e., the relative importance, of each input feature to assess the added value of the PBHM's predictions on the final model forecasts. To summarize, the main research questions addressed in this study are as follows:

1. How does the overall forecast accuracy of the LSTM models compare to that of ARIMA, particularly for longer lead times? In this regard, it will be interesting to see whether the nonlinear LSTM outperforms the linear ARIMA model when using the same input data, and to what extent the LSTM can leverage additional meteorologic inputs.

2. Can the LSTM models achieve a higher forecast accuracy than ARIMA for flood events? Notably, particularly in these instances, a high forecast accuracy is crucial in operational flood forecasting settings.

3. Is the PBHM-HLSTM able to extract valuable information from the underperforming PBHM's results?

2 Study area and data

In this study, we investigated one medium-sized catchment located in the foothills of the Austrian Alps. The catchment drains an area of about 78 km^2 and features elevations from approximately 600 to 1600 m above sea level. The catchment features one gauging station operated by the Hydrographic Service of Styria (Austria). The mean annual discharge at the gauging station is approximately 1.0 $m^3 s^{-1}$. The largest flood events in the catchment mostly occur during the summer months at a sub-daily timescale. Figure 1 provides an overview of the catchment's geographic location, its boundaries, the position of the gauging station, and the river network.

The catchment presented here was part of a broader study in which multiple catchments were modeled using a conceptual rainfall–runoff model (Gegenleithner et al., 2024a). Specifically, Gegenleithner et al. (2024a) employed the distributed wflow_hbv model (Schellekens et al., 2020). Due to the characteristics of the catchments investigated, the model was set up with a temporal resolution of 15 min. For most of the catchments presented in Gegenleithner et al. (2024a), the rainfall–runoff model displayed a high model accuracy, with Nash–Sutcliffe efficiency (NSE) values of 0.77 or higher and Kling–Gupta efficiency (KGE) values of 0.83 or higher. However, for the catchment presented in this study, the model demonstrated notably poorer performance. For the studied period (2011–2017), it merely achieved an NSE of 0.43, a KGE of 0.74, and a percent bias (PBIAS) of +16.0. For a detailed explanation of these performance metrics, refer to Appendix B. Additionally, the PBHM displayed significant shortcomings in capturing the flood hydrograph characteristics, i.e., the rising and falling limbs of the hydrographs as well as the timing and magnitude of the peak discharge.

To develop our forecast models, we utilized the results of the PBHM at the gauge's location (see Fig. 1), denoted as Q_{sim}. Additionally, we incorporated the observed discharge, henceforth referred to as Q_{obs}. For the PBHM-HLSTM exclusively, we also included meteorologic forcings as an input. Specifically, 1 km × 1 km rasters of total precipitation and near-surface temperature, obtained from the Integrated Nowcasting through Comprehensive Analysis (INCA) system (Haiden et al., 2011) and provided by GeoSphere Austria, were utilized. From the raster data, we extracted the catchment's mean and maximum precipitation, designated as p_{mean} and p_{max}, respectively, along with its mean temperature, T_{mean}. All datasets were available at 15 min intervals. A comprehensive overview of the used data and their key statistics is provided in Table A1.

3 Methodology

3.1 Development of the forecast models

To conduct this study, we developed a total of three model variants. The first model, ARIMA, relied on forecasting the residuals between the simulated and observed discharge. Subsequently, the forecasted residuals were used to correct the PBHM's forecasts. The second model, eLSTM, was based on a hindcast–forecast LSTM network, which (similar to ARIMA) used simulated and observed discharge to obtain the forecasts. However, contrary to ARIMA, the LSTM model directly predicted the discharge in the forecast period. The third model, PBHM-HLSTM, was developed with the same architecture as the eLSTM, but it was supplied with additional meteorologic input, namely the mean and maximum catchment precipitation as well as its mean temperature.

Considering the nature of the catchment investigated, all forecast models were developed with a temporal resolution of 15 min and a 24 h forecast horizon, equivalent to 96 consecutive forecast steps.

3.1.1 Model optimization: time series cross-validation

To optimize the hyperparameters of our ARIMA and LSTM models, we employed a blocked cross-validation strategy, as recommended by Bergmeir and Benítez (2012). Furthermore, we chose an expanding-window setup, which allowed us to evaluate the model performance on a multitude of previously unseen data by progressively expanding the data available for training, validation, and testing. Especially in hydrologic modeling applications, where the data exhibit considerable variability (e.g., dry vs. wet years), this strategy can boost the model's performance on unseen data.

We implemented our cross-validation strategy by initially dividing the available time series into equally sized folds, i.e., subsets of the data. Each fold consisted of a sample size of $N = 34\,903$, approximately equivalent to 1 year's worth of data. This procedure resulted in seven folds, corresponding to the years 2011 through 2017. Subsequently, we utilized these folds to create a total of five cross-folds used for model training, validation, and testing. Following the expanding-window strategy, each cross-fold was extended by one fold compared with the previous one. Within each cross-fold, the last and second-to-last folds served as the testing and validation sets, while all preceding folds were used for model training.

For optimizing the models, we employed two loops. In the inner loop, the parameters of each model were optimized using the training and validation sets of each cross-fold. Following the recommendations of Tashman (2000), the models underwent retraining for each cross-fold. For the LSTM models, the hyperparameters were tuned in the outer loop. Thereby, the performance of multiple candidate models was evaluated for the test sets, and the one that minimized the tuner objective function was chosen for final deployment. For the objective function, we selected a combination of the NSE and KGE metrics. For a detailed description of the employed objective function, refer to Appendix C. For ARIMA, the hyperparameters were defined by evaluating the PBHM's residuals, the overall model performance, and ARIMA's model residuals. However, similar to the LSTM models, ARIMA's model parameters were fitted on the training sets of the cross-folds. A visual representation of the methodology presented here is provided in Fig. 2.

3.1.2 Autoregressive integrated moving average (ARIMA) model

ARIMA-type models are widely used for predicting hydrometeorological time series, such as precipitation or discharge (Brath et al., 2002; Broersen and Weerts, 2005; Liu et al., 2015; Khazaeiathar et al., 2022). ARIMA models are commonly denoted as ARIMA(p, d, q), where p is the order of the autoregressive part, d is the differentiation order, and q represents the order of the moving average component. In other words, the values of p and q indicate the number of previous values considered for making the forecasts, and d specifies the number of differentiation operations applied to the original time series. The ARIMA model presented here relies on forecasting the residuals of the PBHM's simulated discharge and that observed at the gauging station, i.e., $e = Q_{\text{sim}} - Q_{\text{obs}}$. The forecasted residuals, $\hat{e}$, are then used to correct the PBHM's forecasts. A visual representation of this procedure is given in Fig. 3.

The ARIMA model presented in this study was developed using the Python "statsmodels" library (Seabold and Perktold, 2010). We assumed the PBHM's residuals to be approximately Gaussian. Furthermore, we assumed that the PBHM's residuals are correlated, stationary (or can be made stationary by ARIMA), and preferably close to homoscedastic. On closer inspection, we found that the residuals exhibited a high degree of heteroscedasticity, which could be stabilized by applying a Box–Cox transformation (Box and Cox,

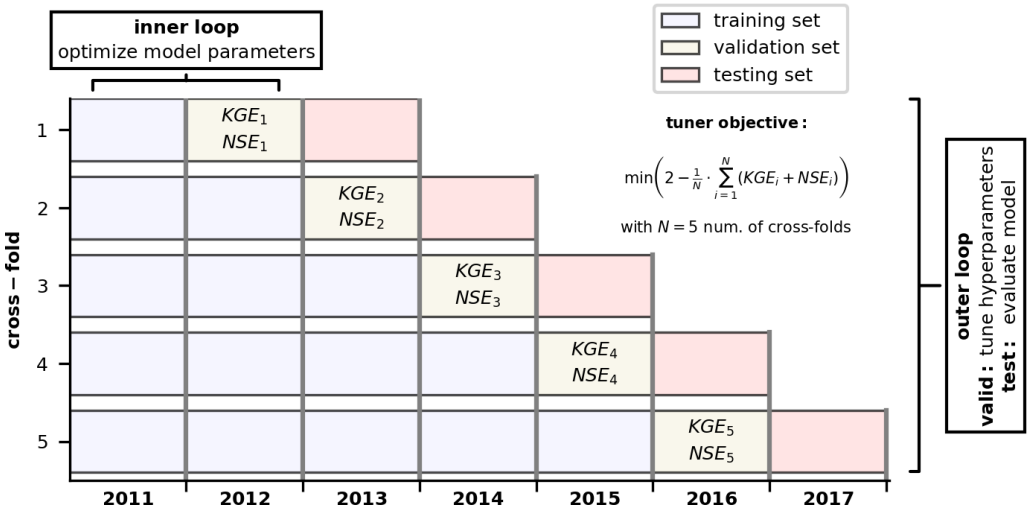

Figure 2. Blocked cross-validation strategy with an expanding-window setup. The parameters of the models were fitted within the inner loop, while the hyperparameters were tuned in the outer loop, utilizing the validation fold of each of the five cross-folds.

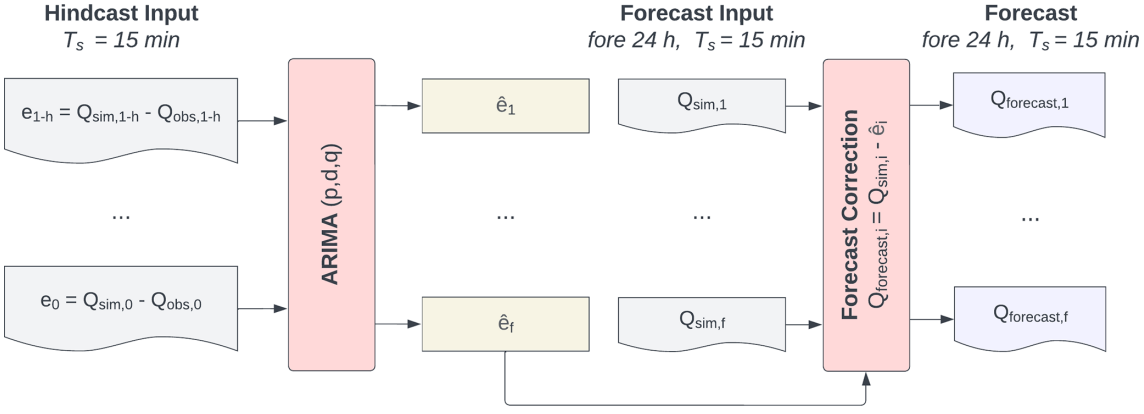

Figure 3. The ARIMA architecture. The optimized ARIMA(p, d, q) model utilized the residuals between the PBHM's results (Q_{sim}) and the observed discharge (Q_{obs}) in the past (e) to forecast the residuals in the forecast period ($\hat{e}$). Consecutively, the forecasted residuals were used to correct Q_{sim} in the forecast period. The terms h and f refer to the hindcast and forecast periods, respectively.

1964) to the PBHM's results and the observed discharge prior to computing the residuals, as shown by studies such as Li et al. (2021). A fixed λ value of 0.2 was used for the Box–Cox transformation, which has proven itself in hydrologic model applications (e.g., Li et al., 2021; Engeland et al., 2010). The Box–Cox transformation also improved Gaussianity. For a detailed statistical evaluation of the residuals, refer to Appendix A2. Stationarity was checked by investigating the autocorrelation function (ACF), which showed a slow decay over many lags, typically indicating some degree of nonstationarity (see Fig. A2a). To make the time series stationary, we added one differentiation operation to the ARIMA model ($d = 1$), which was found to be sufficient for the data used in this study. In addition to the ACF, we also computed the partial autocorrelation function (PACF; see Fig. A2b). The ACF and PACF were then used to get a first estimate of the q and p orders of the ARIMA model. Considering the narrow 5 % significance bounds and the rather low correlations, we iteratively determined the optimum model orders by evaluating ARIMA's overall model performance in the testing folds, whilst not overfitting the model. Additionally, we evaluated ARIMA's model residuals, which ideally should be independent, homoscedastic, and normally distributed. First, ARIMA's model residuals displayed some remaining correlation structures. Second, we also found that the residuals displayed some degree of non-Gaussianity and to a lesser degree heteroscedasticity, independent of the model configuration used. To summarize, the optimum model configuration for the ARIMA model presented in this study was ARIMA$(5, 1, 6)$.

Contrary to other studies (e.g., Broersen and Weerts, 2005), our ARIMA model was not retrained adaptively, i.e., in each forecast step. Instead, ARIMA's model coefficients were determined by utilizing the entire training time series

of each cross-fold (see Sect. 3.1.1), and the resulting coefficients were used for the forecasts in the validation and test sets. Notably, a comparable approach was also employed by Brath et al. (2002).

3.1.3 Hindcast–forecast long short-term memory network (PBHM-HLSTM and eLSTM)

Long short-term memory networks (Hochreiter and Schmidhuber, 1997) are a special form of recurrent neural networks (RNNs). They are specifically designed to address the common issue of vanishing gradients that are often encountered during the training process of RNNs. RNNs process sequential data by maintaining hidden states (H) that retain information from previous inputs, allowing them to capture temporal dependencies. In addition, LSTM networks possess cell states (C) and incorporate three gates – namely, the input gate for controlling incoming information to the cell state, the output gate for regulating the passage of information to the hidden state, and the forget gate for determining the retention or clearance of stored information in the cell state.

The LSTM models presented in this study were developed using TensorFlow (Abadi et al., 2015) and the Keras framework (Chollet, 2015). Both LSTM variants were implemented with a hindcast–forecast architecture, similar to the one presented by Gauch et al. (2021) and Nevo et al. (2022). This architecture involved coupling two distinct LSTM layers, one for the hindcast period and one for the forecast period. The sequence-to-one hindcast LSTM learned patterns in the data of the past. Subsequently, the hindcast LSTM's last hidden H_0 and cell states C_0 were extracted and handed to a fully connected layer. The output of this layer was then used to initialize the first hidden H_1 and cell states C_1 of the sequence-to-sequence forecast LSTM. Besides information on the hindcast period, which was given by the states of the hindcast LSTM, the forecast LSTM included additional features available in the forecast period. The sequential output of the forecast LSTM was then flattened and passed through another fully connected layer to obtain the discharge forecasts for the next 24 h. For this layer, we used the rectified linear unit (ReLU) as the activation function, which prevented negative discharge forecasts.

To prevent data leakage, the models' input features were normalized based on statistics calculated from the first available year (2011). For the normalization, we used min–max scaling for the discharge and precipitation data, while z score standardization was used for the temperature. The models were trained using the mean-squared error (MSE) as a loss function. The hyperparameter tuning was conducted by employing the Adam optimizer (Kingma and Ba, 2015), which minimized a combined objective function consisting of the KGE and NSE metrics (see Appendix C).

The architecture presented in Fig. 4 was used to develop two model variants. The first variant, eLSTM, solely included the observed discharge in the hindcast as well as the simulated discharge in both the hindcast and forecast periods. The second model variant, PBHM-HLSTM, additionally included meteorologic forcings; specifically, the catchment's mean and maximum precipitation as well as its mean temperature in both the hindcast and forecast periods were used. Additionally, PBHM-HLSTM incorporated discharge observations in the hindcast period and the PBHM's results in both the hindcast and forecast periods, respectively.

To optimize the models' hyperparameters, we employed a random grid search tuner (O'Malley et al., 2019) as the outer loop of the cross-validation strategy presented in Sect. 3.1.1. Auxiliary information on the parameters subjected to optimization as well as the models' final hyperparameters can be found in Appendix C.

3.1.4 Sensitivity analysis of neural networks – integrated gradients

To assess the importance of each input feature processed by the LSTM model, we used the integrated gradients (IG) method (Sundararajan et al., 2017). This evaluation was exclusively conducted for the PBHM-HLSTM, which included all input features used in this study. The integrated gradients were evaluated for the model's output, which can be written as follows:

$$\mathrm{IG}_i^{\mathrm{approx}}(\boldsymbol{x}) = \left(x_i - x_i'\right) \times \sum_{k=1}^{m} \frac{\partial F\left(\boldsymbol{x'} + \frac{k}{m}(\boldsymbol{x} - \boldsymbol{x'})\right)}{\partial x_i} \times \frac{1}{m}, \quad (1)$$

where $\boldsymbol{x}$ is the input of interest; F is the model; $\boldsymbol{x'}$ is the baseline (in our case, a sequence of zeros, as suggested by Kratzert et al., 2019a); x_i is the input in the ith dimension, i.e., at the ith input node; and m is the step size of the approximation of the integral (here, 50, as suggested by Sundararajan et al., 2017). In our case, the output of the model is a sequence of size 96, representing the forecast steps. The number of input dimensions, i.e., input nodes, accumulates from five hindcast features (each with a sequence of size 48) and four forecast features (each with a sequence of size 96), resulting in 624 integrated gradients per output node and sample.

3.2 Model performance evaluation

We utilized the five cross-folds (2013 through 2017) presented in Sect. 3.1.1, specifically the test sets, to evaluate the performance of our forecast models. In alignment with the research questions addressed in this study, we conducted the following evaluations:

- *How does the overall forecast accuracy of the LSTM models compare to that of ARIMA, particularly for longer lead times?* To answer this question, we first evaluated each model's (ARIMA, eLSTM, and PBHM-HLSTM) annual performance, i.e., the overall performance for each of the 5 previously unseen testing years.

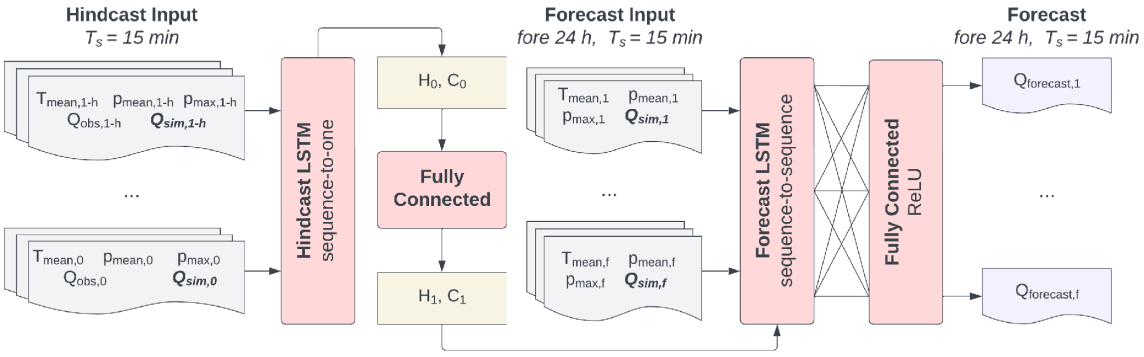

Figure 4. The LSTM architecture. The optimized LSTM models incorporated the PBHM's simulations (Q_{sim}) in both the hindcast and forecast periods as well as the observed discharge at the gauging station (Q_{obs}). The PBHM-HLSTM exclusively incorporated the meteorologic quantities p_{mean}, p_{max}, and T_{mean} in both the hindcast and forecast periods. The hidden and cell states of the hindcast LSTM (H_0 and C_0) were used to initialize the hidden and cell states of the forecast LSTM (H_1 and C_1). The terms h and f refer to the hindcast and forecast periods, respectively.

For this evaluation, we utilized well-established metrics in hydrology, namely the NSE, the KGE, and the PBIAS. Additionally, we included the FHV high flow bias, which evaluates the model bias for the highest 2 % of the flow duration curve. The formulation of the FHV can be found in Appendix B, alongside the other performance metrics. For each metric, we computed the annual average across the 24 h forecast horizon as well as the individual values corresponding to the 96 forecast steps. Moreover, we also monitored the propagation of the mean absolute error (MAE) and the variability in the absolute errors (AE) for each lead time step. The variability was assessed by computing the standard deviation of the absolute errors for each forecast step. In general, models with a high forecast accuracy are expected to display an MAE close to zero and a low standard deviation.

– *Can the LSTM models achieve a higher forecast accuracy than ARIMA for flood events?* This question was addressed by conducting a detailed investigation of each model's performance for the two largest flood events in each year. Specifically, we evaluated how well the models were able to capture the maximum peak discharge with respect to both the timing and magnitude. To measure this, we computed each model's median timing error ($e_{\Delta t}$) as well as the median peak magnitude error (e_{peak}) across all forecasts in a predefined evaluation window. The timing error quantifies the median temporal offset between the maximum observed and simulated peak discharge in number of time steps. Similarly, the magnitude error quantifies the median difference between the maximum observed and simulated peak discharge in percent. To add to this, we also evaluated the distribution of the MAE and the variability in the absolute errors. This was done analogously to the methodology presented in the previous point, although for the highest 2 % of the discharge values only.

– *Is the PBHM-HLSTM able to extract valuable information from the underperforming PBHM's results?* This question was addressed by evaluating the importance of each input feature by employing the IG method presented in Sect. 3.1.4. In accordance with the previous research questions, we evaluated the PBHM-HLSTM's overall feature importance as well as the importance specifically for flood events, again for the two largest flood events per year. The overall importance was assessed by calculating the integrated gradients from the sum of all values at the output nodes and was evaluated for all testing folds. On the other hand, the feature importance for the flood events was determined by computing the IG from the maximum value at the output nodes, which was evaluated for all samples when the maximum peak was present in the forecast horizon. This approach enabled us to assess the importance of each feature at different distances from the maximum discharge peak. For instance, how important is the observed discharge when the maximum peak is three steps away from the forecast origin, t_0.

4 Results

4.1 Overall model performance

4.1.1 Annual average model performance

Evaluating the average annual model performance for the NSE, KGE, PBIAS, and FHV metrics showed that all model variants improved upon the underperforming PBHM's results. Each model's annual performance metrics, averaged over the 24 h forecast horizon, are reported in Table 1.

Table 1. Average annual model performance comparison. Shown are the KGE, NSE, PBIAS, and FHV efficiency metrics. All metrics are averaged over the entire forecast horizon and are reported for each testing year. The best values per metric and year are highlighted in bold.

Year	PBHM				ARIMA				eLSTM				PBHM-HLSTM			
	KGE	NSE	PBIAS	FHV	KGE	NSE	PBIAS	FHV	KGE	NSE	PBIAS	FHV	KGE	NSE	PBIAS	FHV
2013	0.63	0.19	+15.8	+32.4	**0.90**	0.79	**+0.4**	**−0.7**	0.89	0.85	−1.1	+2.3	0.87	**0.92**	−4.7	+1.5
2014	0.74	0.49	+15.6	+6.9	0.89	0.79	**+0.7**	**+5.8**	0.85	0.90	+2.1	−13.4	**0.94**	**0.95**	−3.1	−8.8
2015	0.51	0.24	+6.2	+35.5	0.82	0.70	**+0.6**	+22.4	**0.88**	**0.90**	+1.3	−5.4	0.87	0.88	+7.7	**+4.7**
2016	0.74	0.51	+10.5	**−5.2**	0.84	0.70	**+0.3**	**−5.2**	0.68	0.68	−8.6	−38.6	**0.89**	**0.88**	−2.6	−9.8
2017	0.19	−4.24	+60.0	+74.5	0.60	0.22	**+0.5**	+7.4	0.79	0.64	−0.7	**−1.0**	**0.83**	**0.70**	+12.2	+10.1

The results revealed that the LSTM-based models excelled in terms of NSE and KGE, and this was found to be especially true for the PBHM-HLSTM. For instance, the PBHM-HLSTM was able to achieve an average NSE value of 0.92 in 2013 compared with the 0.19 of the original PBHM. Even in the worst-performing year, 2017, the PBHM-HLSTM was able to elevate the average KGE and NSE values of the PBHM from 0.19 and −4.24 to 0.83 and 0.70, respectively. Overall, the PBHM-HLSTM was found to outperform both ARIMA and the eLSTM in terms of the NSE and KGE in most of the years evaluated, and in the years where this was not the case, the differences in performance were marginal. A different image is drawn when investigating the models' bias metrics (PBIAS and FHV). Particularly in terms of PBIAS, ARIMA was able to outperform the other model variants. We found that this was due to ARIMA's high performance in cases where the forecasts followed a clear pattern or trend, which is often the case under baseflow conditions in hydrologic model applications. Although ARIMA also showed a comparably high performance for the FHV bias, the performance gap to the LSTM models was less distinct. In fact, the PBHM-HLSTM showed the most consistent results in this regard, producing no significant outliers.

The significant performance gap between the PBIAS and the NSE and KGE metrics, however, suggested shortcomings in the forecasts obtained by ARIMA. The most straightforward way to identify these shortcomings was by dissecting the individual components of the KGE efficiency metric. This metric consists of three components that measure the linear correlation, the bias, and the variability between the simulated and observed discharge. As expected, the KGE's bias term for the ARIMA forecasts was close to perfect. Furthermore, the variability term did not signal systematic shortcomings compared to the results of the LSTM networks. However, regarding the linear correlation term, we found that the LSTM forecasts significantly outperformed those of ARIMA. According to Gupta et al. (2009), this term is majorly influenced by the model's ability to capture the peak timing as well as the rising and falling limbs of the hydrographs.

4.1.2 Average model performance over lead time

Each model's performance was also assessed by monitoring the development of the NSE, KGE, PBIAS, and FHV metrics across the 24 h forecast horizon (96 consecutive time steps). The results of each testing year and metric are presented in Fig. 5.

As anticipated, both the ARIMA and the LSTM forecasts surpassed the PBHM's results across most evaluated metrics and years. ARIMA, in particular, performed well in terms of both bias metrics (PBIAS and FHV). The only exception was found to be ARIMA's high FHV in 2015. Moreover, in terms of the NSE and KGE, ARIMA's forecast accuracy was comparably high for the first couple of forecast steps. However, this initial accuracy was shown to quickly decline with increasing lead time. This became particularly evident in 2017, when ARIMA's initial KGE dropped from 0.98 in the first prediction step to 0.60 in the last. An even more significant performance decrease was observed for the NSE metric, for which ARIMA achieved an initial value of 0.97 in the first step but 0.29 in the last. Compared with ARIMA, the LSTM models displayed a different forecast behavior. First, the bias metrics of both LSTM models were mostly higher when compared with those obtained by ARIMA, particularly the PBIAS. Interestingly, when solely judged by their bias metrics, both LSTM variants suggested more or less equal model performance, outperforming each other in some of the years used for evaluation. Arguably, the PBHM-HLSTM achieved more consistent forecasts, considering that the eLSTM produced a significant FHV bias in 2016. Second, the LSTM networks consistently performed worse than ARIMA in the first forecast steps, as suggested by both the KGE and NSE metrics. However, contrary to ARIMA, they tended to mostly sustain their initial accuracy across the forecast horizon. This was found to be most pronounced for the NSE, although it was also observed to a lesser degree for the KGE. For instance, even in the worst-performing year, 2017, the PBHM-HLSTM was able to uphold an NSE of 0.62 and a KGE of 0.73 across the 24 h forecast horizon. Comparing the eLSTM and PBHM-HLSTM model variants, the latter clearly showed superior model performance when judged by the NSE and KGE metrics. Besides a few exceptions where both models performed on par, the PBHM-HLSTM outper-

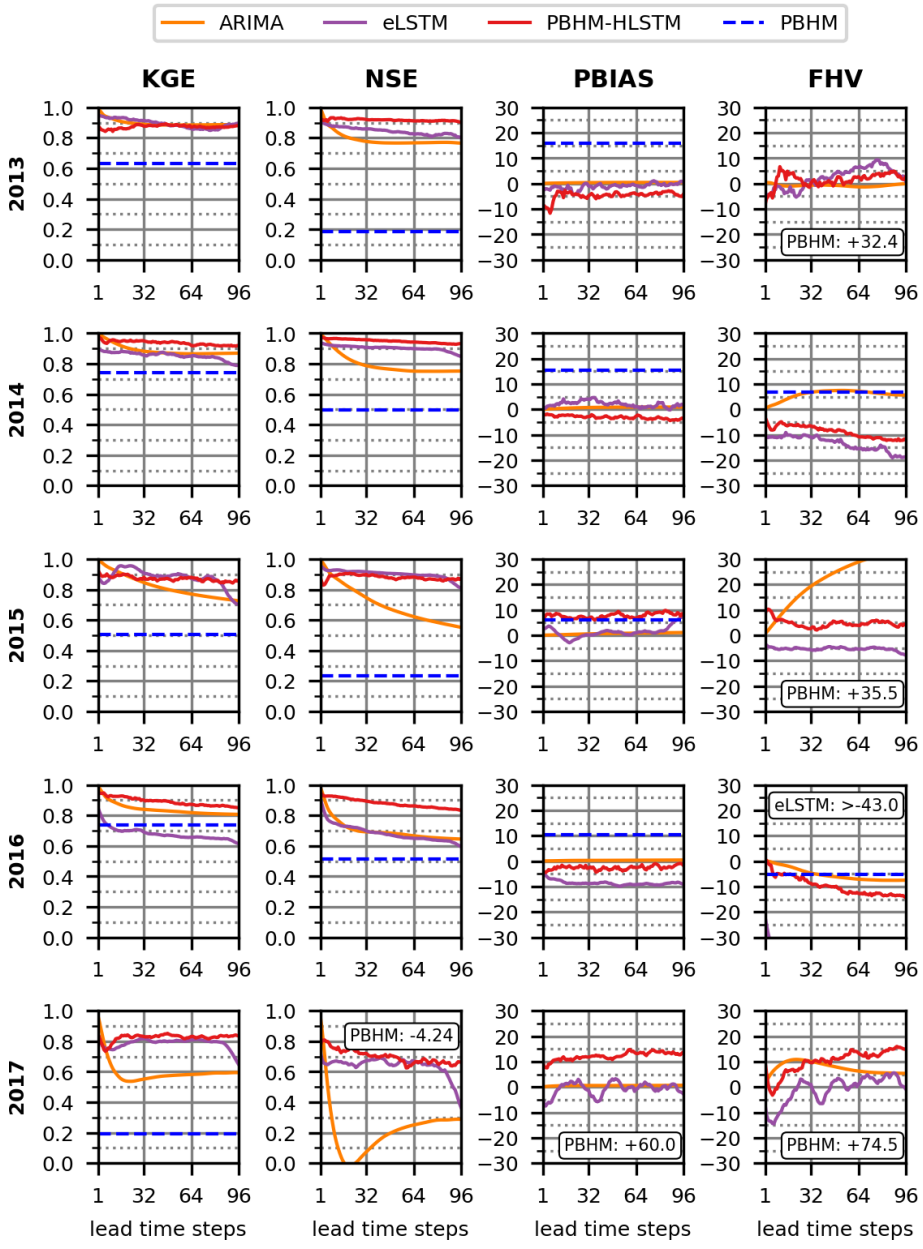

Figure 5. Development of the KGE, NSE, PBIAS, and FHV metrics over the 24 h (96 lead time steps) forecast horizon. Included are all of the developed model variants and all testing years.

formed the eLSTM across all years used for evaluation in this regard. This clearly highlights the additional benefit of adding meteorologic forcings into model development.

In addition to the presented metrics used for model evaluation, we also measured the forecast performance by means of the mean absolute error (MAE) and the standard deviation of the absolute errors (σ(AE)). Both measures were evaluated across the forecast horizon and are shown in Fig. 6. Analogously to the results presented in Fig. 5, the trend in the mean absolute error (Fig. 6, left column) displays ARIMA's high forecast accuracy in the first couple of prediction steps. However, it also reaffirmed its gradual decline in accuracy. In contrast, the LSTM model variants displayed larger errors in the first steps, but their decline in forecast accuracy was less pronounced. Interestingly, in some years (i.e., 2013, 2015, and 2017) the MAE of the LSTM-based models was found to be higher throughout the entire forecast horizon. At first glance, this contradicts the results presented in Fig. 5, particularly when focusing on the NSE metric. This apparent contradiction, however, can be explained by the variability in the errors shown in Fig. 6 (right column). Unlike the LSTM-based models, the forecasts generated by ARIMA

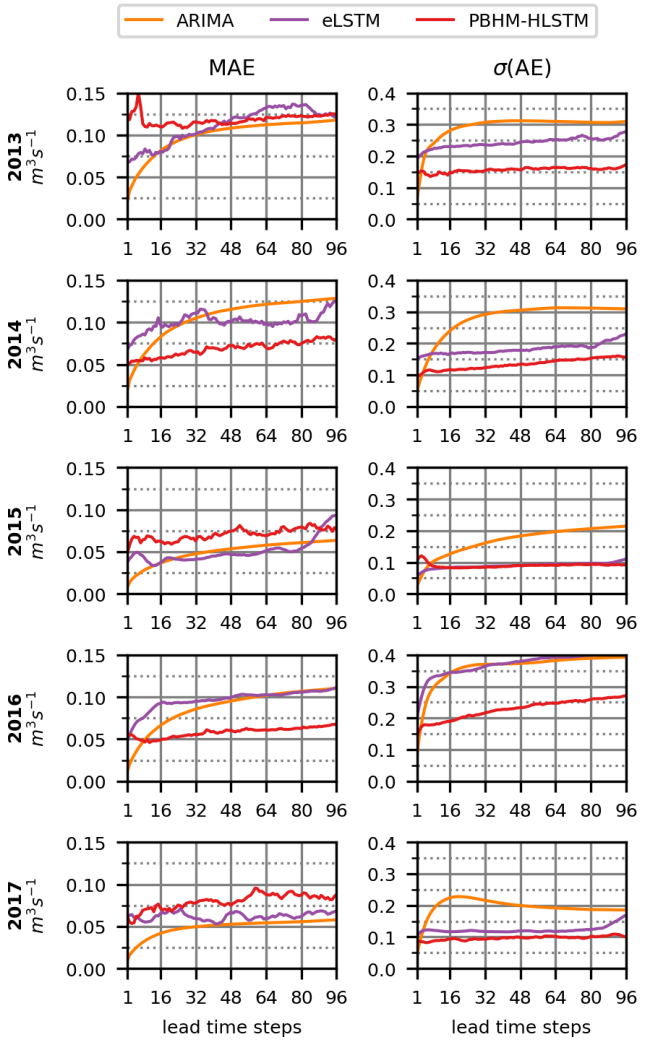

Figure 6. Development of the absolute errors for all flows. Shown are the MAE and the standard deviation (σ) of the AE per testing year for the 24 h forecast horizon (96 time steps).

demonstrated significant variability in their errors. This indicates that, while it produced highly accurate forecasts in some cases, ARIMA often yielded predictions that deviated substantially from the actual outcomes, especially for more temporally distant forecasts. In contrast, both LSTM variants achieved a considerably lower error variance. Quantitatively, the LSTM-based models provided more reliable forecasts on average after three forecast steps (i.e., 45 min).

4.2 Performance for elevated river discharge

4.2.1 Peak timing and magnitude errors

For assessing the performance of our forecast models for flood events, we determined the models' median peak magnitude and timing errors for the two largest flood events in each year. Positive magnitude errors indicate model overestimation, whereas negative values suggest an underestimation. As for the timing errors, negative values indicate that the model predicted the maximum peak discharge earlier than observed, whereas positive values indicate the opposite. The results of this evaluation are presented in Table 2.

Upon initial inspection, the evaluation of the peak magnitude and timing errors reaffirmed the deficiencies of the PBHM in capturing the flood runoff dynamics. Specifically, the substantial timing errors suggest shortcomings of the model with respect to adequately depicting the characteristics of the flood event hydrographs. In terms of magnitude error, the PBHM achieved a median value of −49.9 %, predominately underestimating the observed peak discharge. Arguably, none of the investigated model variants were able to precisely pinpoint the magnitude of the flood events. However, by far the best performance in this regard was shown by PBHM-HLSTM, which achieved a median magnitude error of −27.5 %. Interestingly, the worst performance in this regard was shown by the eLSTM model, while ARIMA's results fell between those of the eLSTM and the PBHM-HLSTM. It has to be mentioned that ARIMA, in contrast to eLSTM, generally exhibited a lower magnitude error in the first steps, which improved the overall value reported.

In terms of timing error, the ARIMA-corrected forecasts showed no improvement compared to the PBHM's original forecasts. In contrast, the eLSTM was able to majorly reduce the timing errors in the forecasts. Considering the fact that both models were supplied with the same input data, this clearly shows that the linear correction model (ARIMA) was not able to adequately transform the shape of the poorly depicted hydrographs of the PBHM. Comparing the timing errors of the eLSTM and PBHM-HLSTM reaffirmed the superiority of the PBHM-HLSTM. Specifically, the PBHM-HLSTM displayed a median timing error of merely two time steps across all events, which corresponds to 30 min in this study. In contrast, the median timing error of the eLSTM was found to be five time steps. Auxiliary information on the models' predictions for the largest flood events in each year can be found in Appendix D.

4.2.2 Performance over lead time for elevated river discharge

Analogously to the results presented in Fig. 6, we evaluated the development of the mean absolute error (MAE) and the standard deviation of the absolute errors ($\sigma(AE)$) across the forecast horizon, although only for the largest 2 % of the discharge values. The results of this investigation are shown in Fig. 7.

While ARIMA often outperformed or matched the MAE of the LSTM networks when considering all flows (see Fig. 6, left column), the evaluation of the largest discharge values clearly demonstrates the superiority of the LSTM-based models (see Fig. 7, left column). Particularly the PBHM-HLSTM, although to a lesser degree also the eLSTM,

Table 2. Comparison of the median peak magnitude error (e_{peak}, in percent) and timing error ($e_{\Delta t}$, in number of time steps) for the two largest flood events in each year. The smallest errors and offsets per event are highlighted in bold.

Year	Event	Obs. peak discharge	PBHM		ARIMA		eLSTM		PBHM-HLSTM	
		($m^3 s^{-1}$)	e_{peak} (%)	$e_{\Delta t}$	e_{peak} (%)	$e_{\Delta t}$	e_{peak} (%)	$e_{\Delta t}$	e_{peak} (%)	$e_{\Delta t}$
2013	First	15.00	−90.3	31	−71.7	47	−78.0	59	**−38.4**	**2**
	Second	10.02	**+13.3**	20	−26.1	19	−40.0	**−1**	−27.5	−5
2014	First	7.27	**+3.5**	18	+4.5	19	−27.0	**0**	−27.8	−2
	Second	6.23	+23.3	16	**+19.1**	16	−24.8	**2**	−26.5	4
2015	First	5.85	−62.5	36	−29.9	36	−66.7	10	**−25.2**	**2**
	Second	3.33	**+4.7**	49	+17.4	44	−32.7	**0**	−13.4	3
2016	First	17.94	−73.6	26	**−45.9**	29	−77.1	5	−68.8	**3**
	Second	9.99	−45.4	**18**	−56.7	20	−65.5	27	**−43.9**	68
2017	First	9.21	−49.9	25	−48.9	48	−54.5	3	**−17.3**	**−1**
	Second	7.37	−63.1	28	−32.6	31	−59.9	6	**+8.4**	**1**
All folds	First		−62.5	26	−45.9	36	−66.7	5	**−27.8**	**2**
	Second		**+4.7**	20	−26.1	20	−40.0	**2**	−26.5	3
	Both		−49.9	26	−32.6	31	−59.9	5	**−27.5**	**2**

was able to achieve a considerably lower MAE compared with ARIMA. A similar picture was drawn by the variance (as reflected by the standard deviation) in the absolute errors (see Fig. 7, right column), for which the PBHM-HLSTM displayed considerably lower values than both ARIMA and the eLSTM.

4.3 Sensitivity analysis of the PBHM-HLSTM

4.3.1 Overall sensitivity

The average importance of each of the PBHM-HLSTM's input features for both the hindcast and forecast LSTM networks is shown in Fig. 8. As anticipated, the PBHM-HLSTM heavily relied on past discharge observations (O_{obs}) for deriving its forecasts. Interestingly, the importance of the observations seemed to decay exponentially with increasing distance to the forecast origin (t_0). As shown in Fig. 8, the influence of the observations almost dampened out after approximately 48 time steps. This means that the model gave increasingly more weight to observations close to t_0. Furthermore, in the annual average, the model seemed to rely very little on past and future precipitation, p_{max} and p_{mean}, most likely because both variables were zero or close to zero throughout most of the year. In comparison, the mean temperature (T_{mean}) in both the hindcast and forecast had some influence on the predictions, most likely adding seasonality context to the model. The PBHM's simulated discharge was found to have the second-highest impact on the forecasts, right after the observations. Particularly in the forecast period, the simulated discharge influenced the final predictions considerably.

Table 3 summarizes the normalized feature importance values for all evaluation years, averaged over the hindcast and forecast periods, respectively. This evaluation reaffirmed that the model highly valued the observed discharge, which was found to be the most important feature for all years. Moreover, the high importance of the PBHM's simulated discharge was consistent across all years, surpassed only by the observations. Surprisingly, in 2017, the simulated discharge had the highest relative importance of all years, although it featured the worst performance of the PBHM.

4.3.2 Sensitivity for flood events

To investigate the importance of the individual features for flood events, we exclusively evaluated the integrated gradients for the two largest flood events of each year. Figure 9 shows the importance of the hindcast (panel a) and forecast (panel b) features for various distances of the predicted discharge peak to the forecast origin (t_0). In this regard, one means that the predicted peak is located at t_{0+1} and analogously at t_{0+96} for a value of 96.

The results show that the closer the peak was located to the forecast origin, the more the forecast was influenced by the observed discharge. This comes as no surprise, as the observed discharge at t_0 should be a reasonable predictor of the discharge at t_{0+1}. Interestingly, also in the case where the peak was located at the end of the forecast horizon t_{0+96}, the observed discharge still had a rather high impact on the forecast. As for the precipitation features (p_{mean} and p_{max}), the importance of the former was found to be considerably higher. This means that the model gained more information from the precipitation volume than from its intensity. Fur-

Table 3. Normalized feature importance values per testing year. The values were normalized by the total sum of importance values per year. The most important input feature per year is highlighted in bold. Values less than or equal to 0.01 are omitted to increase readability.

Year	Hindcast features					Forecast features			
	Q_{sim}	p_{max}	T_{mean}	p_{mean}	Q_{obs}	Q_{sim}	p_{max}	T_{mean}	p_{mean}
2013	0.09		0.03		**0.55**	0.19		0.08	0.04
2014	0.08		0.04	0.02	**0.55**	0.18		0.08	0.04
2015	0.08		0.08		**0.51**	0.16		0.12	0.03
2016	0.07		0.07	0.02	**0.49**	0.18		0.10	0.04
2017	0.06		0.12		**0.32**	0.24	0.02	0.16	0.07
All folds	0.08		0.06		**0.51**	0.19		0.10	0.04

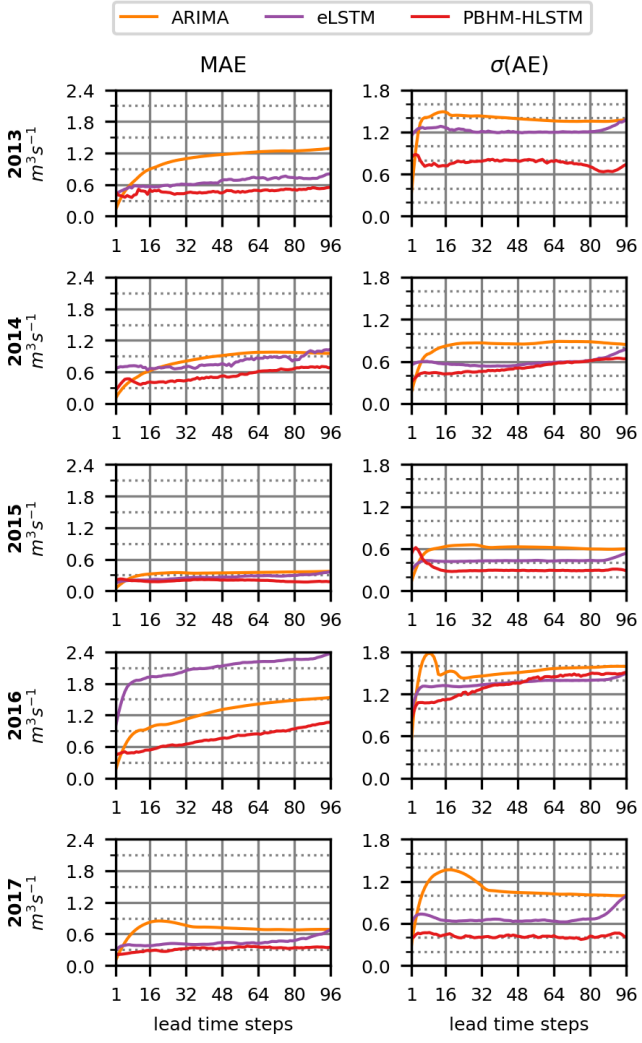

Figure 7. Development of the absolute errors for the largest 2 % of the discharge values. Shown are the MAE and the standard deviation (σ) of the AE per year for the 24 h forecast horizon (96 time steps).

ther investigating the mean precipitation feature revealed additional insights. First, the hindcast p_{mean} was shown to have a high influence when the peak was close to t_0, which then decayed exponentially with increasing distance of the peak discharge to the forecast origin. From a theoretical point of view, this makes sense, as some of the precipitation at this point has already passed the gauging station as surface runoff. Second, the forecast p_{mean} showed little importance for forecasts that were close to the forecast origin, but its importance was shown to grow rapidly with increasing distance to t_0. This occurs as the rainfall needs time to concentrate and does not directly result in runoff. Moreover, for predictions for flood events, it was shown that the PBHM-HLSTM relied on the PBHM's output. In the hindcast, it was found to be the second-most important feature, while in the forecast, its importance was found to be generally equal to that of the maximum precipitation (p_{max}) and the mean temperature (T_{mean}).

Table 4 summarizes the normalized feature importance values for the two largest flood events per year, averaged over the hindcast and forecast periods, respectively. The results clearly show that the mean precipitation in the forecast period had the highest relative importance of all input features. In fact, it showed that the model mostly relied on forecast features for its predictions during flood events, with the only exception being the observed discharge in the hindcast. As for the Q_{sim}, p_{max}, and T_{mean} features in the forecast, they were all shown to have a more or less equal influence on the final flood forecasts.

5 Discussion

In this study, we built upon the promising outcomes of prior research (see Rozos et al., 2021; Konapala et al., 2020; Frame et al., 2021) by exploring the potential of LSTM networks to enhance the forecast accuracy of PBHMs employed in operational flood forecasting systems. Following the approaches of Gauch et al. (2021) and Nevo et al. (2022), we developed our LSTM models using a hindcast–forecast architecture. This architecture was chosen as it facilitates an effective

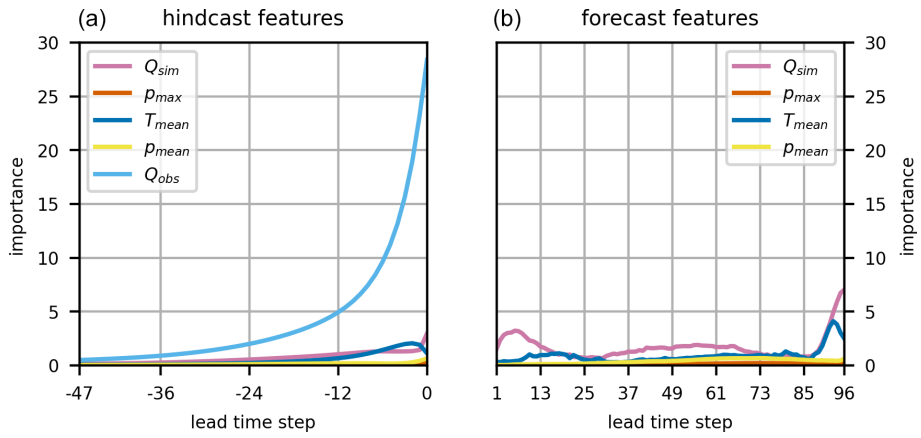

Figure 8. Importance of all input features summed over all testing years. Shown are the feature importance values of the hindcast features **(a)** and the forecast features **(b)**.

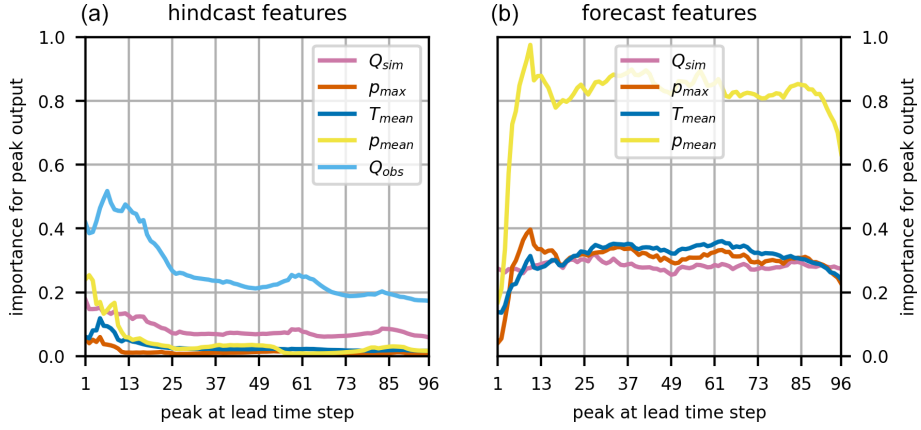

Figure 9. Importance of input features for the peak prediction in the forecast window summed over the two largest flood events per year. Shown are the feature importance values of the hindcast features **(a)** and the forecast features **(b)**.

integration into operational forecasting systems. Specifically, the hindcast–forecast architecture allows for a clear separation between hindcast and forecast data, which comes with certain advantages. For example, this strategy would allow the model to distinguish between meteorologic forecasts and analyses, potentially enabling it to learn from their differences. Furthermore, the cross-validation strategy presented here enables a seamless, continuous improvement of the model as new data become available. To assess the benefits of the LSTM-based forecasts, we developed two LSTM model variants and compared their forecast skill to that of a conventional ARIMA model, using one underperforming PBHM as a case study. To ensure comparability between the LSTM and ARIMA approaches, one LSTM (eLSTM) was restricted to use the same data as ARIMA, whereas the other incorporated additional meteorologic variables (PBHM-HLSTM). Of particular interest was how the LSTM approach improved prediction accuracy, especially for flood events and for longer lead times – both being recognized weaknesses of ARIMA for cases in which the underlying PBHM provides poor initial estimates.

When comparing the forecasts obtained by the LSTM and ARIMA models, we observed that both methods had certain advantages and disadvantages. ARIMA generally demonstrated a very high accuracy in the first forecast steps. However, this initial accuracy was often shown to quickly decline with increasing lead time. These findings align with those presented in previous studies, such as Brath et al. (2002) or Broersen and Weerts (2005). In contrast, the LSTM networks generally exhibited a larger error in the first steps but were able to mostly sustain their initial accuracy over the 24 h forecast horizon. This became particularly evident when observing the variance in the absolute errors. Both LSTM models, particularly the PBHM-HLSTM, displayed a considerably lower error variance compared with the results obtained by ARIMA. This suggests that, in comparison to ARIMA, they produced notably fewer poor forecasts. Interestingly, ARIMA performed particularly well in terms of PBIAS. The

Table 4. Importance of features for the peak prediction in the forecast window. The values were normalized by the total sum of importance values per event. The most important input feature per event is highlighted in bold. Values less than or equal to 0.01 are omitted to increase readability.

Year	Event	Hindcast features					Forecast features			
		Q_{sim}	p_{max}	T_{mean}	p_{mean}	Q_{obs}	Q_{sim}	p_{max}	T_{mean}	p_{mean}
2013	First	0.02		0.02		0.09	0.04	0.11	0.11	**0.59**
	Second	0.12		0.02	0.04	0.20	**0.28**	0.02	0.12	0.19
2014	First	0.06		0.02	0.02	**0.32**	0.22	0.02	0.11	0.23
	Second	0.07		0.02	0.06	0.20	**0.24**	0.06	0.15	0.21
2015	First			0.03	0.03	0.05	0.05	0.12	0.18	**0.52**
	Second	0.02				0.11	0.16	0.11	0.22	**0.37**
2016	First	0.02				0.14	0.10	0.18	0.15	**0.36**
	Second	0.12			0.03	0.21	0.24	0.02	0.10	**0.26**
2017	First					0.02	0.07	0.27	0.16	**0.44**
	Second						0.03	0.24	0.15	**0.54**
All folds	First	0.02				0.12	0.10	0.17	0.14	**0.41**
	Second	0.06			0.03	0.13	0.17	0.10	0.15	**0.34**
	Both	0.04			0.02	0.12	0.13	0.14	0.14	**0.38**

reason for that was found in ARIMA's high accuracy for forecasts that followed a clear trend or pattern, which occurs most often under baseflow conditions in hydrologic model applications.

When focusing solely on the forecast skill at specific flood events, the LSTM networks clearly outperformed ARIMA. This became particularly evident when investigating the models' timing errors, i.e., the temporal offset between the maximum observed and simulated peak discharge. While both LSTM variants were able to significantly reduce the initial timing errors of the PBHM, this was not achieved by ARIMA. This implies that ARIMA was not able to adequately transform the event hydrographs in instances where the underlying PBHM was not able to give an adequate initial estimation, a fact that was also shown by Liu et al. (2015). As for the magnitude errors, i.e., the difference between the maximum observed and simulated discharge, only the PBHM-HLSTM was able to achieve somewhat satisfying results. Interestingly, the eLSTM even performed worse than ARIMA in this regard. This indicates that the eLSTM did not receive sufficient context from the observed and simulated discharge alone to accurately capture the magnitude of flood events. This underscores the importance of incorporating meteorologic variables when employing LSTM models in operational forecasting systems. Notably, the ARIMA model could also potentially benefit from the inclusion of meteorologic variables. In a preliminary study, we tested an ARIMAX model that included the catchment's mean precipitation as an exogenous variable. However, for the case presented here, the results of the ARIMAX and ARIMA models were nearly identical and were, thus, excluded from the study.

Considering the comparably high performance of the PBHM-HLSTM in this study and, more generally, the remarkable capabilities of LSTM models in predicting river discharge (e.g., Kratzert et al., 2019b), a question is raised regarding the added benefits that the underperforming PBHM provides. To assess the added value of the PBHM in this study, we evaluated the relative importance of each of the PBHM-HLSTM's input features. Our findings indicate that, on average, the PBHM-HLSTM model heavily relied on the results of the PBHM, particularly its forecasts. In fact, the PBHM's discharge predictions were identified as the second-most important feature, following the observed discharge. While the PBHM-HLSTM specifically for flood events did, to some extent, also rely on the PBHM's forecasts, the mean catchment precipitation emerged as the most important feature in these instances. These findings also explain the large performance gap between the eLSTM and the PBHM-HLSTM for flood events.

When employing forecast-enhancing models in operational flood forecasting systems, several important considerations must be taken into account. First and foremost, such models are not an all-in-one device suitable for every purpose. Although the PBHM-HLSTM presented here was able to significantly improve upon the PBHM's forecasts, it is still a post-processing technique that is meant to enhance predictions at the specific location of the gauging station, while leaving the PBHM's system states untouched. However, often these system states, e.g., the state of the snow cover, the soil moisture, or spatially distributed information

of the runoff, function as an additional decision criterion for the system's operator and are often used for implementing more complex forecasting chains. Considering the poor performance of the PBHM in this study, its system states are most likely not correct and can, thus, not provide any added benefit. Furthermore, it has to be considered that there is a reason why the PBHM's performance is poor. Often this can be linked to poor model parameterization, the inability of the model to capture some important catchment processes, or uncertainties in the input data. For the latter, these uncertainties might be present in the data used for setting up the PBHM, in the meteorologic forcings, or in the data used for calibration (e.g., the gauge discharge). Notably, in contrast to PBHMs, data-driven models (e.g., LSTM networks) might be adept at learning any systematic errors embedded in the data, consequently improving forecast accuracy. Overall, we believe that data-driven forecast-enhancing strategies are highly valuable in contexts like the one presented in this study, where the PBHM alone fails to deliver satisfactory forecasts.

Although the PBHM-HLSTM model presented here has already achieved a comparably high forecast accuracy, potential exists for future enhancements. First, refining the preprocessing phase, especially through more targeted feature engineering could further enhance the model's predictive capabilities. Second, the target data (gauge discharge) can be diagnosed. For instance, the probe technique presented by Lees et al. (2022) could be adopted to identify behavioral anomalies in the LSTM cell states by comparing multiple catchments. Lastly, future work could also focus on investigating a hybrid ARIMA–LSTM approach, potentially further increasing the model's prediction accuracy, particularly in the first forecast steps.

6 Conclusions

In this study, we explored the potential of long short-term memory (LSTM) networks as a post-processing strategy for enhancing the forecast performance of an underperforming process-based hydrologic model (PBHM). We specifically compared this post-processing strategy to a conventional autoregressive integrated moving average (ARIMA) model, as such models are often employed in operational flood forecasting systems. Our focus was on the models' performance for extended lead times and, particularly, for flood events, with both being critical aspects in operational flood forecasting. To facilitate an objective comparison, we developed two LSTM model variants. One variant, eLSTM, was restricted to use the same input data as ARIMA, namely observed discharge and the discharge generated by the PBHM, whereas the other, PBHM-HLSTM, additionally incorporated meteorologic variables. Furthermore, we assessed the added value of the underperforming PBHM's results on the predictions of the PBHM-HLSTM by evaluating the importance of each of the model's input features. The main findings of this study can be summarized as follows:

- All model variants (ARIMA, eLSTM, and PBHM-HLSTM) significantly enhanced the forecast accuracy of the existing PBHM.
- ARIMA achieved a particularly high accuracy in the first forecast steps. However, this initial accuracy declined quickly with increasing lead time. In contrast, the LSTM models showed a larger initial error but mostly maintained their initial accuracy over the 24 h forecast horizon.
- ARIMA showed shortcomings in forecasting the discharge for flood events. Specifically, it failed to accurately predict the timing and the maximum peak discharge of the flood events. The eLSTM improved timing predictions but significantly underestimated the magnitude of the events. Only the PBHM-HLSTM was able to sufficiently predict both the timing and the magnitude of the flood events.
- Despite the PBHM's poor performance, the PBHM-HLSTM still considered its output informative. On an annual average, the PBHM's output was found to be the second-most important feature, following the observed discharge. For flood event predictions the PBHM's results were also found to be important, but the catchment's mean precipitation was identified as the most critical input feature in these cases.

To summarize, in this study, we demonstrated that LSTM models can pose a viable alternative to frequently employed ARIMA correction models in operational flood forecasting systems.

Appendix A: Statistics of the input data

A1 Model input data

Table A1 shows the key statistics of the observed and simulated discharge as well as the meteorologic forcings used in this study.

A2 PBHM model residuals

Two statistical tests have been employed to analyze the PBHM's residuals. First, the goodness-of-fit test was used to analyze how closely the residuals follow a Gaussian distribution. For this purpose, the Filliben r correlation value (Filliben, 1975) was computed, for which a value close to 1 signifies a Gaussian distribution. Second, the Lagrange multiplier statistic of the Breusch–Pagan test (Breusch and Pagan, 1979) was evaluated to assess the degree of heteroscedasticity of the residuals. The critical value for homoscedasticity was computed as 3.84 for a 5 % significance level based on the Chi distribution given 1 degree of freedom. The application of the Box–Cox transform, using a λ value of 0.2, showed an increase in Gaussianity in the residuals' distribution as well as a reduction in heteroscedasticity, even below the critical value for homoscedasticity. Figure A1 shows a quantile–quantile (Q–Q) plot including the Filliben r test statistics for the original and Box–Cox transformed residuals (panel a), alongside a scatterplot of the PBHM's residuals against the observed discharge, which includes the test statistics of the Breusch–Pagan test (panel b).

A3 Autocorrelation evaluation of the PBHM residuals

We evaluated the autocorrelation function (ACF) and partial autocorrelation function (PACF) for the PBHM's residuals. Both are visualized in Fig. A2. The correlation values and their 5 % significance bounds were obtained by bootstrapping, where the residuals were analyzed for each year and the results were averaged. Figure A2 includes the original PBHM residuals, the Box–Cox transformed residuals, and the residuals following one differentiation operation.

Table A1. Statistics of the catchment's runoff (gauge observation, Q_{obs}; PBHM simulation, Q_{sim}) as well as its mean precipitation, maximum precipitation, and temperature (p_{mean}, p_{max}, and T_{mean}, respectively).

			Year						
Parameter	Statistic	Unit	2011	2012	2013	2014	2015	2016	2017
Q_{obs}	μ	$m^3 s^{-1}$	0.57	1.01	1.21	1.17	0.71	0.83	0.57
	σ	$m^3 s^{-1}$	0.25	0.84	0.68	0.67	0.33	0.68	0.23
	max	$m^3 s^{-1}$	9.61	25.20	15.00	7.27	5.85	17.90	9.21
	Σ	hm^3	18.0	31.9	38.1	36.8	22.4	26.2	17.7
Q_{sim}	μ	$m^3 s^{-1}$	0.62	1.10	1.40	1.35	0.76	0.92	0.91
	σ	$m^3 s^{-1}$	0.33	0.83	1.02	0.72	0.52	0.70	0.48
	max	$m^3 s^{-1}$	4.20	8.94	11.40	7.68	4.50	6.43	4.61
	Σ	hm^3	19.5	34.9	44.1	42.6	23.8	29.0	28.3
p_{max}	max	$mm\,h^{-1}$	118	180	100	84.6	109	231	173
p_{mean}	max	$mm\,h^{-1}$	29.2	69.7	45.8	33.5	38.6	61.6	69.3
	Σ	mm	871	1289	1284	1225	912	1188	1153
T_{mean}	μ	°C	6.88	6.72	6.43	7.47	7.56	6.98	6.94
	σ	°C	7.97	8.72	8.21	6.72	7.90	7.59	8.27

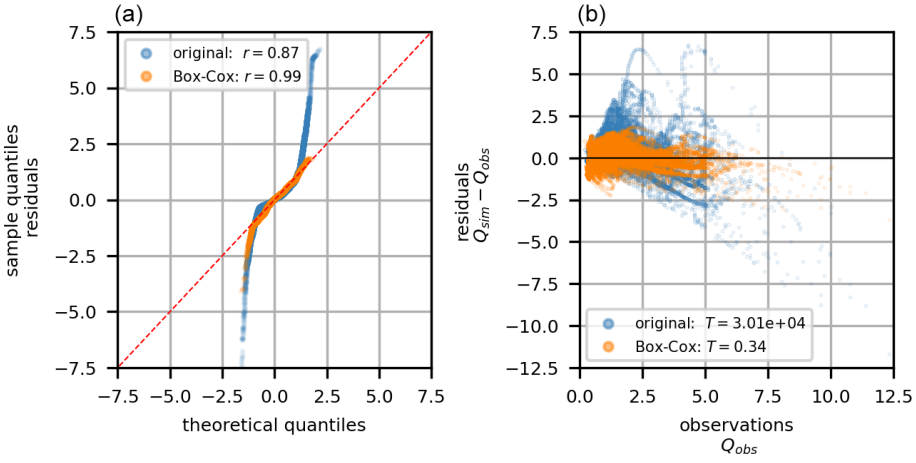

Figure A1. A Q–Q plot with Filliben r test statistics for the original and Box–Cox-transformed residuals. **(a)** Scatterplot of the PBHM's original and transformed residuals against the observed discharge including the test statistics of the Breusch–Pagan test **(b)**.

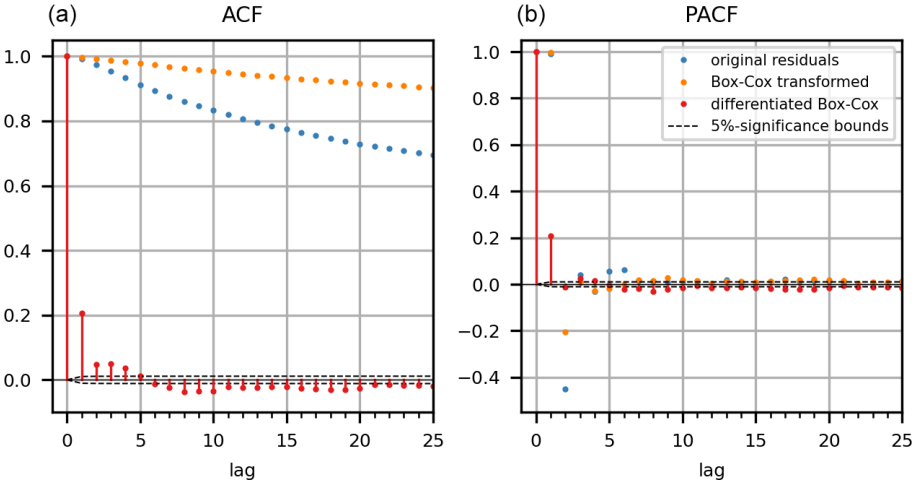

Figure A2. Autocorrelation function (ACF; **a**) and partial autocorrelation function (PACF; **b**) for the original PBHM model residuals, the Box–Cox-transformed residuals, and the residuals following one differentiation operation.

Appendix B: Evaluation metrics

B1 Nash–Sutcliffe efficiency (NSE)

The Nash–Sutcliffe efficiency (NSE; Nash and Sutcliffe, 1970) quantifies how well the model performs compared to a simple mean discharge benchmark. In its original form, the NSE can be written as follows:

$$\mathrm{NSE} = 1 - \frac{\sum_{t=1}^{N}(Q_{\mathrm{obs},t} - Q_{\mathrm{sim},t})^2}{\sum_{t=1}^{N}(Q_{\mathrm{obs},t} - \overline{Q}_{\mathrm{obs}})^2}, \tag{B1}$$

where $Q_{\mathrm{obs},t}$ and $Q_{\mathrm{sim},t}$ are the observed and predicted discharge, respectively. The NSE is bound between 1 and $-\infty$, with 1 indicating perfect model predictions.

B2 Kling–Gupta efficiency (KGE)

The Kling–Gupta efficiency (KGE) was proposed by Gupta et al. (2009). It is a combined efficiency metric that considers the correlation, the bias, and the variability in the flow. In this study, we utilized the modified Kling–Gupta efficiency (Kling et al., 2012), which can be written as follows:

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + (\beta-1)^2 + (\gamma-1)^2}, \tag{B2}$$

where r is the correlation term; β is the bias term given by the ratio of the mean of the simulated and observed discharge values, $\mu_{\mathrm{sim},t}/\mu_{\mathrm{obs},t}$; and γ is the variability term, which is computed from the standard deviations and the mean values as $\frac{\sigma_{\mathrm{sim},t}/\mu_{\mathrm{sim},t}}{\sigma_{\mathrm{obs},t}/\mu_{\mathrm{obs},t}}$. The KGE is bound between 1 and $-\infty$, with 1 indicating perfect model predictions.

B3 Percent bias (PBIAS)

The PBIAS is a measure that quantifies if the model tends to underpredict or overpredict the observed discharge. It can be written as follows (Yilmaz et al., 2008):

$$\mathrm{PBIAS} = \frac{\sum_{t=1}^{N}(Q_{\mathrm{sim},t} - Q_{\mathrm{obs},t})}{\sum_{t=1}^{N} Q_{\mathrm{obs},t}} \cdot 100, \tag{B3}$$

where $Q_{\mathrm{sim},t}$ and $Q_{\mathrm{obs},t}$ are the observed and predicted discharge, respectively. The PBIAS can take both positive and negative values, where positive values indicate that the model on average overpredicts the observations and vice versa. A PBIAS close to 0 indicates a widely unbiased model.

B4 High-segment volume percent bias (FHV)

The FHV quantifies the bias at high flows with an exceedance probability lower than 0.02 based on the flow duration curve (Yilmaz et al., 2008). It can be written as follows:

$$\mathrm{FHV} = \frac{\sum_{i=1}^{H}(Q_{\mathrm{sim},i} - Q_{\mathrm{obs,i}})}{\sum_{i=1}^{H}(Q_{\mathrm{obs,i}})} \cdot 100, \tag{B4}$$

where $Q_{\mathrm{obs,i}}$ and $Q_{\mathrm{sim},i}$ are the observed and predicted discharge, respectively, and $i = 1, 2, \ldots H$ is the index of the flow value located within the high-flow segment of the flow duration curve.

Appendix C: Auxiliary information on LSTM hyperparameter tuning

To tune the hyperparameters, we selected a combined objective function (f_{obj}) consisting of the NSE and KGE metrics. The objective function was computed as follows:

$$f_{\text{obj}} = 2 - \text{KGE} - \text{NSE}, \tag{C1}$$

where 0 would indicate a perfect fit by the model.

Table C1 shows the LSTM hyperparameters subjected to optimization, their search space, and their final values after tuning. Additionally, we investigated two different hindcast lengths, namely 12 and 24 h, and chose the final model variants based on the lowest objective function value.

Figure C1 depicts the training and validation losses per epoch for all five folds for the selected model variants. The tuner used an early-stopping mechanism by monitoring the development of the validation loss.

Table C1. Hyperparameter tuning information, defined search space, and final parameter set for the LSTM models.

Parameter	Search space		eLSTM*	eLSTM-H48	PBHM-HLSTM-H96	PBHM-HLSTM*
	Min	Max				
ID best trial			41	40	48	22
Objective			0.232	0.239	0.169	0.162
No. of LSTM units	4	32	17	22	20	23
Initial learning rate	1×10^{-3}	1×10^{-2}	0.00774	0.0090	0.0065	0.00912
Dropout rate	0.01	0.5	0.247	0.0228	0.0633	0.039
Batch size			4000	4000	4000	4000
Retrain epochs			5	5	5	5
Hindcast length			96	48	96	48

* Selected for further processing based on tuner objective.

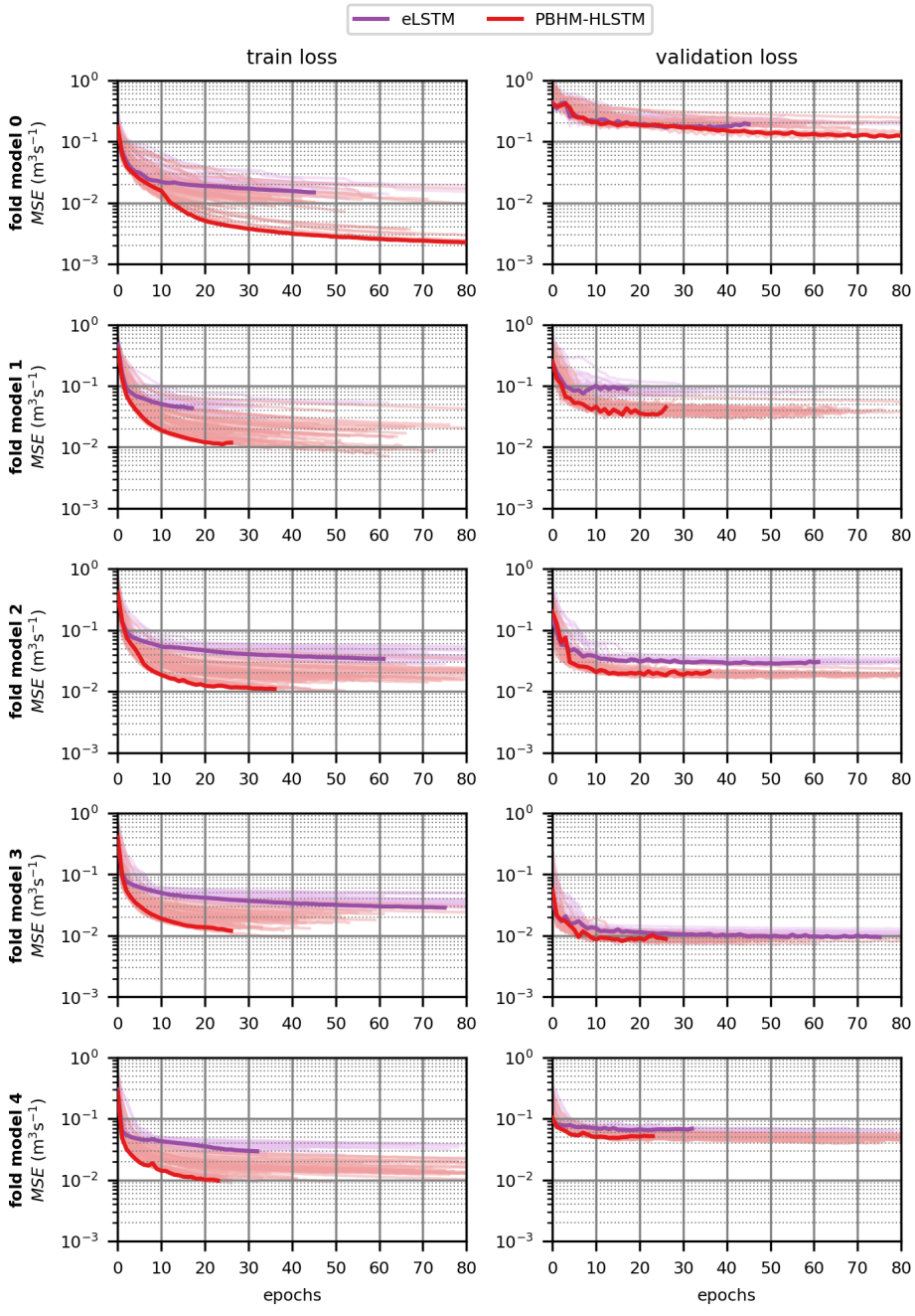

Figure C1. Best models' losses during training and validation.

Appendix D: Model predictions for the largest flood event per year

Figure D1 shows the location and magnitude of the estimated flood peak for all 96 lead time predictions for the largest flood event per year. Cumulative average precipitation over the catchment and the PBHM's predictions are given as a reference. It can be seen that the predicted peaks of the PBHM-LSTM model incorporating information on precipitation and temperature during the forecast horizon matched more closely to the actual peaks than the predictions from the variants solely built on the PBHM's results and the observed discharge. A summary of these findings can also be found in Table 2.

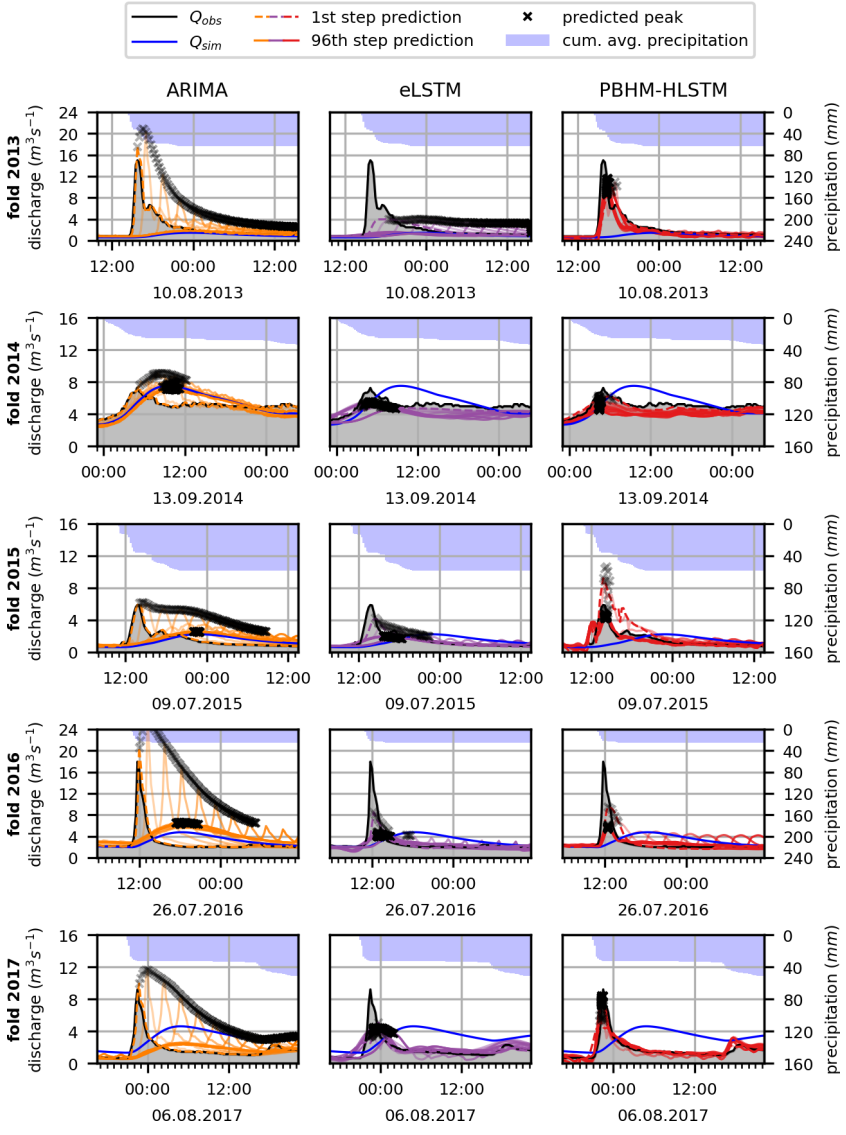

Figure D1. Forecast comparison for the largest flood events per year. Given are the results of the ARIMA, eLSTM, and PBHM-HLSTM models.

Code and data availability. The Python code and processed data presented in this study are stored at https://doi.org/10.5281/zenodo.10907245 (Gegenleithner et al., 2024b). The published data were derived from the following datasets:

1. *Gauge discharge data* were sourced from the Styrian Government, Department 14 – Water Management, Resources and Sustainability (Hydrographic Service of Styria). The data were validated by the provider. The time stamps were converted from GMT+1 to UTC by the authors.

2. *Meteorologic data* were provided by GeoSphere Austria. More specifically, 1 km × 1 km rasters were provided from which we extracted catchment-averaged values. Those averaged values are included in the dataset.

3. *Hydrologic modeling results* were obtained from Gegenleithner et al. (2024a).

Author contributions. SG: conceptualization, methodology, data curation, and writing – original draft preparation. MP: conceptualization, methodology, data curation, and writing – original draft preparation. CD: funding acquisition and writing – review and editing. RK: writing – review and editing. JS: supervision and writing – review and editing.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. We declare that, during the preparation of this work, we used generative AI to enhance specific sections of the written content. The content was reviewed, and we take full responsibility for the quality of this publication.

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. The authors wish to express their gratitude to the Styrian Government, Department 14 – Water Management, Resources and Sustainability (Hydrographic Service of Styria) and to GeoSphere Austria for providing the data for this study. This research was supported by the TU Graz Open Access Publishing Fund.

Review statement. This paper was edited by Ralf Loritz and reviewed by Niels Schuetze and one anonymous referee.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X.: TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, software available from tensorflow.org, https://www.tensorflow.org/ (last access: 15 March 2024), 2015.

Addor, N., Jaun, S., Fundel, F., and Zappa, M.: An operational hydrological ensemble prediction system for the city of Zurich (Switzerland): skill, case studies and scenarios, Hydrol. Earth Syst. Sci., 15, 2327–2347, https://doi.org/10.5194/hess-15-2327-2011, 2011.

Bergmeir, C. and Benítez, J.: On the use of cross-validation for time series predictor evaluation, Inform. Sciences, 191, 192–213, https://doi.org/10.1016/j.ins.2011.12.028, 2012.

Borsch, S., Simonov, Y., Khristoforov, A., Semenova, N., Koliy, V., Ryseva, E., Krovotyntsev, V., and Derugina, V.: Russian rivers streamflow forecasting using hydrograph extrapolation method, Hydrology, 9, 1, https://doi.org/10.3390/hydrology9010001, 2021.

Box, G. and Cox, D.: An Analysis of Transformations, J. R. Stat. Soc. B, 26, 211–252, https://doi.org/10.1111/j.2517-6161.1964.tb00553.x, 1964.

Brath, A., Montanari, A., and Toth, E.: Neural networks and non-parametric methods for improving real-time flood forecasting through conceptual hydrological models, Hydrol. Earth Syst. Sci., 6, 627–639, https://doi.org/10.5194/hess-6-627-2002, 2002.

Breusch, T. S. and Pagan, A. R.: A Simple Test for Heteroscedasticity and Random Coefficient Variation, Econometrica, 47, 1287–1294, https://doi.org/10.2307/1911963, 1979.

Broersen, P. M. and Weerts, A. H.: Automatic error correction of rainfall-runoff models in flood forecasting systems, in: 2005 IEEE Instrumentationand Measurement Technology Conference Proceedings, 16–19 May 2005, Ottawa, vol. 2, IEEE, 963–968, https://doi.org/10.1109/IMTC.2005.1604281, 2005.

Chollet, F.: Keras, GitHub [code], https://github.com/keras-team/keras (last access: 15 March 2024), 2015.

Engeland, K., Renard, B., Steinsland, I., and Kolberg, S.: Evaluation of statistical models for forecast errors from the HBV model, J. Hydrol., 384, 142–155, https://doi.org/10.1016/j.jhydrol.2010.01.018, 2010.

Filliben, J. J.: The probability plot correlation coefficient test for normality, Technometrics, 17, 111–117, https://doi.org/10.1080/00401706.1975.10489279, 1975.

Frame, J. M., Kratzert, F., Raney, A., Rahman, M., Salas, F. R., and Nearing, G. S.: Post-processing the national water model with long short-term memory networks for streamflow predictions and model diagnostics, J. Am. Water Resour. As., 57, 885–905, 2021.

Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J., and Hochreiter, S.: Rainfall–runoff prediction at multiple timescales with a single Long Short-Term Memory network, Hydrol. Earth Syst. Sci., 25, 2045–2062, https://doi.org/10.5194/hess-25-2045-2021, 2021.

Gegenleithner, S., Krebs, G., Dorfmann, C., and Schneider, J.: Enhancing flood event predictions: Multi-objective calibration using gauge and satellite data, J. Hydrol., 632, 130879, https://doi.org/10.1016/j.jhydrol.2024.130879, 2024a.

Gegenleithner, S., Pirker, M., Dorfmann, C., Kern, R., and Schneider, J.: Supplement to: Long Short-Term Memory Networks for Enhancing Real-time Flood Forecasts: A Case Study for an Underperforming Hydrologic Model, Zenodo [data set], https://doi.org/10.5281/zenodo.10907245, 2024b.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, J. Hydrol., 377, 80–91, 2009.

Haiden, T., Kann, A., Wittmann, C., Pistotnik, G., Bica, B., and Gruber, C.: The Integrated Nowcasting through Comprehensive Analysis (INCA) system and its validation over the Eastern Alpine region, Weather Forecast., 26, 166–183, 2011.

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

Kalman, R. E.: A New Approach to Linear Filtering and Prediction Problems, J. Basic Eng.-T. ASME, 82, 35–45, https://doi.org/10.1115/1.3662552, 1960.

Khazaeiathar, M., Hadizadeh, R., Fathollahzadeh Attar, N., and Schmalz, B.: Daily Streamflow Time Series Modeling by Using a Periodic Autoregressive Model (ARMA) Based on Fuzzy Clustering, Water, 14, 3932, https://doi.org/10.3390/w14233932, 2022.

Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, Proceedings of the 3rd International Conference on Learning Representations (ICLR), 7–9 May 2015, https://doi.org/10.48550/arXiv.1412.6980, 2015.

Kling, H., Fuchs, M., and Paulin, M.: Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios, J. Hydrol., 424–425, 264–277, https://doi.org/10.1016/j.jhydrol.2012.01.011, 2012.

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., Herrnegger, M., Klotz, D., Hochreiter, S., and Klambauer, G.: NeuralHydrology – Interpreting LSTMs in Hydrology, in: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, edited by: Samek, W., Montavon, G., Vedaldi, Hansen, A., Kai, L., and Müller, K.-R., Springer International Publishing, https://doi.org/10.1007/978-3-030-28954-6_19, pp. 347–362, 2019a.

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, 2019b.

Land Kärnten: Austria 10 m Digital Elevation Model, Land Kärnten [data set], https://www.data.gv.at/katalog/dataset/land-ktn_digitales-gelandemodell-dgm-osterreich (last access: 22 September 2022), 2019.

Lees, T., Reece, S., Kratzert, F., Klotz, D., Gauch, M., De Bruijn, J., Kumar Sahu, R., Greve, P., Slater, L., and Dadson, S. J.: Hydrological concept formation inside long short-term memory (LSTM) networks, Hydrol. Earth Syst. Sci., 26, 3079–3101, https://doi.org/10.5194/hess-26-3079-2022, 2022.

Li, D., Marshall, L., Liang, Z., Sharma, A., and Zhou, Y.: Characterizing distributed hydrological model residual errors using a probabilistic long short-term memory network, J. Hydrol., 603, 126888, https://doi.org/10.1016/j.jhydrol.2021.126888, 2021.

Liu, J., Wang, J., Pan, S., Tang, K., Li, C., and Han, D.: A real-time flood forecasting system with dual updating of the NWP rainfall and the river flow, Nat. Hazards, 77, 1161–1182, 2015.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I—A discussion of principles, J. Hydrol., 10, 282–290, 1970.

Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., Hassidim, A., Klotz, D., Kratzert, F., Metzger, A., Nevo, S., Pappenberger, F., Prudhomme, C., Shalev, G., Shenzis, S., Tekalign, T., Weitzner, D., and Matias, Y.: Global prediction of extreme floods in ungauged watersheds, Nature, 627, 559–563, https://doi.org/10.1038/s41586-024-07145-1, 2024.

Nester, T., Komma, J., and Blöschl, G.: Real time flood forecasting in the Upper Danube basin, J. Hydrol. Hydromech., 64, 404–414, 2016.

Nevo, S., Morin, E., Gerzi Rosenthal, A., Metzger, A., Barshai, C., Weitzner, D., Voloshin, D., Kratzert, F., Elidan, G., Dror, G., Begelman, G., Nearing, G., Shalev, G., Noga, H., Shavitt, I., Yuklea, L., Royz, M., Giladi, N., Peled Levi, N., Reich, O., Gilon, O., Maor, R., Timnat, S., Shechter, T., Anisimov, V., Gigi, Y., Levin, Y., Moshe, Z., Ben-Haim, Z., Hassidim, A., and Matias, Y.: Flood forecasting with machine learning models in an operational framework, Hydrol. Earth Syst. Sci., 26, 4013–4032, https://doi.org/10.5194/hess-26-4013-2022, 2022.

O'Malley, T., Bursztein, E., Long, J., Chollet, F., Jin, H., and Invernizzi, L.: Keras Tuner, GitHub [code], https://github.com/keras-team/keras-tuner (last access: 15 March 2024), 2019.

Rozos, E., Dimitriadis, P., and Bellos, V.: Machine learning in assessing the performance of hydrological models, Hydrology, 9, 5, https://doi.org/10.3390/hydrology9010005, 2021.

Schellekens, J., van Verseveld, W., Visser, M., Winsemius, H., Euser, T., Bouaziz, L., Thiange, C., de Vries, S., Boisgontier, H., Eilander, D., Tollenaar, D., Weerts, A., Baart, F., Hazenberg, P., Lutz, A., ten Velden, C., Jansen, M., and Benedict, I.: openstreams/wflow, Zenodo [code], https://doi.org/10.5281/zenodo.593510, 2020.

Seabold, S. and Perktold, J.: statsmodels: Econometric and statistical modeling with python, in: 9th Python in Science Conference, 28 June–3 July 2010, Austin, 57–61, https://doi.org/10.25080/Majora-92bf1922-011, 2010.

Sundararajan, M., Taly, A., and Yan, Q.: Axiomatic Attribution for Deep Networks, in: International Conference on Machine Learning, 6–11 August 2017, Sydney, https://api.semanticscholar.org/CorpusID:16747630 (last access: 15 March 2024), 2017.

Tashman, L. J.: Out-of-sample tests of forecasting accuracy: an analysis and review, Int. J. Forecasting, 16, 437–450, https://doi.org/10.1016/S0169-2070(00)00065-0, 2000.

Umweltbundesamt GmbH: Austrian river network, v17, Umweltbundesamt GmbH [data set], https://www.data.gv.at/katalog/dataset/c2287ccb-f44c-48cd-bf7c-ac107b771246 (last access: 22 September 2022), 2022.

Weerts, A. H. and El Serafy, G. Y.: Particle filtering and ensemble Kalman filtering for state updating with hydrologi-

cal conceptual rainfall-runoff models, Water Resour. Res., 42, https://doi.org/10.1029/2005WR004093, 2006.

Werner, M., Cranston, M., Harrison, T., Whitfield, D., and Schellekens, J.: Recent developments in operational flood forecasting in England, Wales and Scotland, Meteorol. Appl., 16, 13–22, 2009.

Yaghmaei, N., van Loenhout, J., Below, R., and Guha-Sapir, D.: Human cost of disasters, An overview of the last 20 years: 2000–2019, CRED and UNDRR, https://www.undrr.org/publication/human-cost-disasters-overview-last-20-years-2000-2019 (last access: March 2024), p. 30, 2020.

Yilmaz, K. K., Gupta, H. V., and Wagener, T.: A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, Water Resour. Res., 44, https://doi.org/10.1029/2007WR006716, 2008.