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]Graz University of Technology, Institute of Hydraulic Engineering and Water Resources Management, Stremayrgasse 10/II, 8010 Graz, Austria
[2]flow engineering, Lessingstraße 30, 8010 Graz, Austria
[3]Graz University of Technology, Institute of Interactive Systems and Data Science, 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)

Abstract. Flood forecasting systems play a key role in mitigating socio-economic damages caused by ~~flooding~~ flood events. The majority of these systems rely on process-based hydrologic models (~~PBHM~~PBHMs), which are used to predict future river runoff. ~~To enhance the forecast accuracy of these models, many~~ Many operational flood forecasting systems ~~implement error correction techniques, which is particularly important if the underlying hydrologic model is underperforming. Especially,~~ 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, especially AutoRegressive Integrated Moving Average (ARIMA) type models are frequently employed~~for this purpose~~. Despite their high popularity ~~, numerous~~ in flood forecasting, studies have pointed out potential shortcomings of ~~these~~ ARIMA-type models, such as a decline in forecast accuracy with increasing lead time. ~~To overcome the limitations presented by conventional ARIMA models, we propose a novel forecast correction technique based on a hindcast-forecast~~ In this study, we investigate the potential of Long Short-Term Memory (LSTM) ~~network. We showcase the effectiveness of~~ 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 runoff, while the other additionally incorporated meteorologic forcings (PBHM-HLSTM). Considering the ~~proposed approach by rigorously comparing its capabilities to those of an ARIMA model, utilizing one underperforming PBHM as a case study. Additionally, we test whether the LSTM benefits from the~~ PBHM's ~~results or if a similar accuracy can be reached by employing a standalone LSTM. Our investigations show that the proposed LSTM model significantly improves~~ poor performance, we further evaluated if the PBHM-HLSTM was able to extract valuable information from the PBHM's ~~forecasts. Compared~~ results by analyzing the relative importance of each input feature. Contrary to ARIMA, the ~~LSTM achieves a higher~~ LSTMs were able to mostly sustain a high forecast accuracy for longer lead times. ~~In terms of~~ Furthermore, the PBHM-HLSTM also achieved a high prediction accuracy at flood event runoff, which was not the case for ARIMA and the eLSTM. Our results also revealed that the PBHM-HLSTM to some degree relied on the ~~LSTM performs mostly on par with ARIMA in predicting the magnitude of the events. However, the LSTM majorly outperforms ARIMA in~~

~~accurately predicting the timing of the peak runoff. Furthermore, our results provide no reliable evidence of whether the LSTM is able to extract information from the~~ PBHM's results, ~~given the widely equal performanceof the proposed and standalone LSTM models.~~

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 runoff, empowering decision-makers and emergency forces to implement 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 ~~Nevo et al. (2022)~~Nearing et al. (2024) .

To date, most operational flood forecasting systems are built around process-based hydrologic models (PBHM). These models predict future river runoff 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 networks (LSTM, Hochreiter and Schmidhuber, 1997) have gained recognition for their capabilities in accurately modeling river runoff. For example, Kratzert et al. (2019b) demonstrated that their LSTM model was able to outperform two PBHMs across multiple gauged but also ungauged catchments. Although data-driven models have proven to be a viable alternative to PBHMs for modeling river runoff, they are yet 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 ~~forecasting~~ predicting a sequence of future runoff 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 runoff forecasts based on ~~meteorological~~ 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 runoff). 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 ~~for correcting~~ to enhance the model's forecasts. Considering the available literature, the most relevant ~~correction~~ forecast-enhancing strategies can be grouped as follows: (I) State updating~~(Data Assimilation)~~: 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 ~~or Particle Filters (e.g., Weerts and El Serafy, 2006)~~(Kalman, 1960) or also Particle Filters, as demonstrated by Weerts and El Serafy (2006). (II) Error correction: These methods use observations of one or multiple state variables, mostly river runoff, to ~~correct~~ enhance the hydrologic model's forecasts in a post-processing step. Especially, 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, while for further ahead predictions only slight differences were found to forecasts corrected with the mean runoff over the last three weeks. Secondly, ARIMA models struggle to provide accurate forecasts for flood event runoff when the underlying hydrologic model fails to give an adequate initial estimation, as for example shown by Liu et al. (2015). In their study, Liu et al. (2015) assessed the predictive skills 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 (RNN), to enhance the results obtained from PBHMs, and the outcomes have been remarkably successful. ~~For example~~Although the focus of their study was on model diagnostics, Rozos et al. (2021) demonstrated that ~~the predictive capability of an underperforming PBHM could be improved by employing both a simple RNN and an LSTM~~RNNs and LSTMs, trained on ~~meteorological data as well as~~ 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 also 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 ~~the~~ runoff 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 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 ~~propose a novel~~ developed two LSTM model variants, both implemented with a hindcast-forecast ~~LSTM correction approach and compare its forecast accuracy~~ architecture, similar to the one presented by Gauch et al. (2021) or Nevo et al. (2022), and compared their forecast performances to that of a ~~more~~ conventional ARIMA model~~, using one underperforming PBHM as a case study. Specifically,~~. To ensure comparability, one LSTM variant, eLSTM,

was restricted to use the same input data as ARIMA, specifically observed runoff and that obtained by the PBHM. The second variant, PBHM-HLSTM, was implemented with the same architecture as the ~~selected PBHMhas displayed weaknesses in predicting flood event runoff, i.e., the hydrograph's rising and falling limbs as well as the magnitude and timing of the maximum peak runoff. Besides comparing the efficiencies of ARIMA and the LSTM in correcting the PBHM's forecasts~~eLSTM but additionally incorporated meteorologic forcings. It has to be mentioned that our ARIMA model relied on forecasting residuals, while both LSTM variants directly predicted future runoff. For the PBHM-HLSTM exclusively, we also ~~test an LSTM variant that does not incorporate information from the PBHM. This investigation tests whether the LSTM can extract additional information from the~~ evaluated the contribution, or in other terms, the relative importance, of each input feature to assess the added value of the PBHM's ~~results or if a similar accuracy can be reached by replacing the underperforming PBHM with a standalone LSTM, a question raised by Frame et al. (2021)~~predictions on the final model forecasts. To summarize, the main research questions addressed in this study can be stated as follows: (I) How does the ~~LSTM approach improve the overall quality of the forecasts~~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 non-linear LSTM outperforms the linear ARIMA model when using the same input data, and to what extent the LSTM can leverage additional meteorologic inputs. (II) ~~How does the LSTM approach improve the forecast quality for flood events?~~ Can the LSTM models achieve a higher forecast accuracy than ARIMA at flood event runoff? Notably, high forecast accuracy at flood event runoff is crucial in operational flood forecasting settings. (III) ~~Does the inclusion of the~~ Is the PBHM-HLSTM able to extract valuable information from the underperforming PBHM's results~~improve the predictive skills of the LSTM~~?

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 meters above sea level. The catchment features one gauging station operated by the Hydrographic Service of Styria (Austria). The mean annual runoff at the gauging station is approximately 1.0 $\mathrm{m}^3\mathrm{s}^{-1}$. ~~Flood event runoff in this catchment is primarily driven by heavy precipitation events, with most events occurring~~ The largest flood events in the catchment mostly occur during the summer months at a sub-daily time scale. ~~Fig.~~Figure 1 provides an overview of the catchment's geographic location, its boundaries, the position of the gauging station, as well as the river network.

The ~~here presented catchment~~ catchment presented here was part of a broader study in which ~~various~~ multiple catchments were modeled using a conceptual rainfall-runoff model (Gegenleithner et al., 2024a). ~~This particular catchment was selected for our investigation because the existing~~ Specifically, Gegenleithner et al. (2024a) employed the distributed wflow-hbv model (Schellekens, 2012). Due to the characteristics of the catchments investigated, the model was setup with a temporal resolution of 15 minutes. For most of the catchments presented in Gegenleithner et al. (2024a), the rainfall-runoff model ~~failed to provide accurate runoff predictions.~~ displayed a high model accuracy with Nash-Sutcliffe Efficiencies (NSE) of 0.77 or higher and Kling-Gupta Efficiencies (KGE) of 0.83 or higher. However, for the catchment presented in this study, the model demonstrated

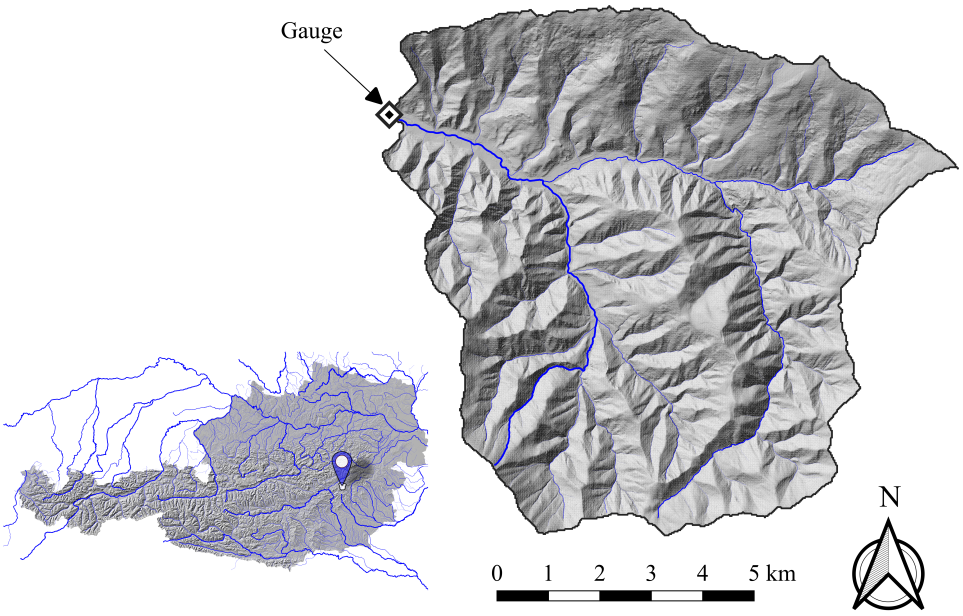


Figure 1. (Bottom left) Location of the study catchment in Austria. (Right) Outline of the study catchment (black line) including the gauging station (black and white diamond) and the main river network (blue lines). This figure was created using the following datasets: Umweltbundesamt GmbH (2022) and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

a notably poorer performance. For the studied period (2011 - 2017) ~~, the existing model merely achieved a Nash-Sutcliffe efficiency (NSE) it merely achieved an~~ NSE of 0.43, a ~~Kling-Gupta Efficiency (KGE)~~ KGE of 0.74, and a Percent Bias (PBIAS) of ~~-16.0+~~16.0. For a detailed explanation of these performance metrics, refer to Appendix B. Additionally, the PBHM displayed significant shortcomings in capturing the flood event runoff characteristics, i.e., the rising and falling limbs of the hydrographs as well as the timing and magnitude of the maximum peak runoffs.

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 ~~discharge measurements~~the observed discharge, henceforth referred to as Q_{obs}. For the ~~LSTM models~~ PBHM-HLSTM exclusively, we also included ~~meteorological~~ meteorologic forcings as an input. Specifically, 1x1 km rasters of total precipitation and near-surface temperature, obtained from the Integrated Nowcasting through Comprehensive Analysis system (INCA, 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}, along with its mean temperature ~~t_{mean}~~, T_{mean}. Noteworthy, all ~~processed~~ datasets were available in 15-minute intervals. ~~An~~ A comprehensive overview of the used data and its key statistics is provided in Table A1.

Statistics of the catchment's runoff (gauge observation Q_{obs}, PBHM simulation Q_{sim}) and meteorological precipitation and temperature forcings (p_{mean}, p_{max} and t_{mean}) comprising of their mean (μ), standard deviation (σ), maximum (max) and annual sum (Σ). **parameterstatisticunit2011201220132014201520162017**$Q_{obs}$$\mu$m^{3}s^{-1}0.571.011.211.170.710.830.57σm^{3}s^{-1}0.250.840

3 Methodology

3.1 Development of the forecast models

For conducting this study, we developed two forecast models, both of which integrated results obtained by the PBHMa total of three model variants. The first model, ARIMA, relied on forecasting the errors residuals between the simulated and observed runoffsrunoff. Subsequently, these errors the forecasted residuals were used to correct the hydrologic modelPBHM's forecasts. The second model, eLSTM, was based on a hindcast-forecast LSTM network. In contrast , which similar to ARIMA used simulated and observed runoff to obtain the forecasts. However, contrary to ARIMA, this the LSTM model directly predicted the runoff by leveraging information on the observed and simulated runoff, along with the meteorological forcings presented in Table A1. Henceforth, we will refer to this modelas HLSTM-PBHM. In addition to these models, we developed a variant of HLSTM-PBHM. This variant was implemented in the forecast period. The third model, PBHM-HLSTM, was developed with the same architecture but without integrating the PBHM's results as a feature. This model will be further referred to as the standalone LSTM or, in short, HLSTMas the eLSTM, but 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 minutes and a 24-hour 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). We Furthermore, we chose an expanding window setup, which allowed us to evaluate the model performances 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 generalization capabilitiesperformance 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 one 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 to 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.

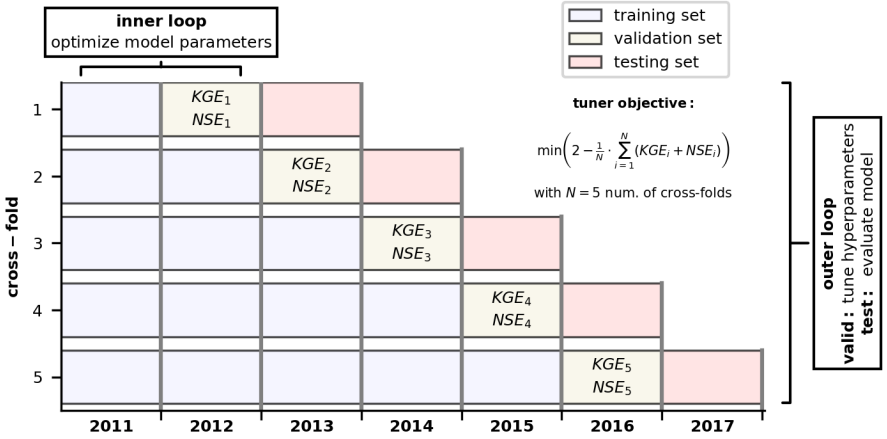


Figure 2. Blocked cross-validation strategy with expanding window setup. The parameters of the models were fitted within the inner loop while the hyperparameters were tuned in the outer loop, utilizing the validation fold of each of the five cross-folds.

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. ~~Subsequently, the models'~~ For the LSTM models, the hyperparameters were tuned in the outer loop. Thereby the performance of multiple candidate models was evaluated for the ~~testing~~ test sets, and the one that ~~maximized~~ 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 as well as 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 here presented methodology is provided in Fig. 2.

3.1.2 AutoRegressive Integrated Moving Average model (ARIMA)

ARIMA-type models are widely used for predicting hydrometeorologic time series ~~,~~ such as precipitation or runoff (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 here presented ARIMA model relies on forecasting the residuals of the PBHM's simulated runoff and that observed at the gauging station, i.e., $e = Q_{sim} - Q_{obs}$. 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 ~~here~~ in this study was developed by using the Python Statsmodels library (Seabold and Perktold, 2010). ~~In the first step the model computed the errors between the gauge observations Q_{obs} and the runoff obtained by~~ 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, 1964) to the PBHM's results and the observed runoff prior to computing the residuals, as for example shown by 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). Noteworthy, the ~~PBHM Q_{sim} in the past. Subsequently, ARIMA predicted the errors in the forecast period and used them to correct the forecasts of the PBHM. A visual representation of this procedure is given in~~ 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 non-stationarity (see Fig. ~~3.~~

~~The here presented ARIMA model was optimized by employing an exhaustive search algorithm, representing the outer loop described in Sect. 3.1.1. The parameters subjected to optimization along with their search space were chosen as follows: $p \in [1, 20]$, $q = p - 1$,~~ A2, left). 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. Additionally to the ACF, we also computed the Partial Autocorrelation Function (PACF, see Fig. A2, right). The ACF and PACF were then used to get a first estimate of the q and ~~$d \in [1, 2]$.~~ 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), ~~the~~ 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 ~~testing sets. In the case presented here, this approach resulted in superior model performances when compared to often employed adaptive model optimization strategies. Noteworthy, similar findings were also presented~~ test sets. Notably, such an approach was also employed by Brath et al. (2002). ~~Following this procedure, the best-performing model was determined as $ARIMA(14, 1, 13)$.~~

3.1.3 Hindcast-forecast Long Short-Term Memory network (~~HLSTM-PBHM~~ PBHM-HLSTM & ~~HLSTM~~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

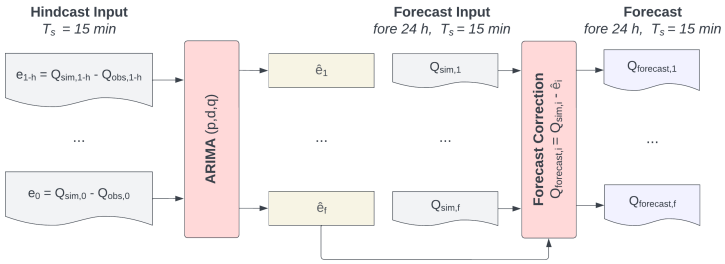


Figure 3. ARIMA architecture. The optimized $ARIMA(p,d,q)$ model utilized the ~~errors~~ residuals between the PBHM's results Q_{sim} and the observed runoff Q_{obs} in the past, e, to forecast the ~~errors~~ residuals in the forecast period $\hat{e}$. Consecutively, the forecasted ~~errors~~ residuals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

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, LSTMs 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 information passage 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 et al., 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, respectively. The sequence-to-one hindcast LSTM learned patterns in the data of the past~~24 hours~~. 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, that 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 runoff forecasts for the next 24 hours. For this layer, we used the Rectified Linear Unit (ReLU) as the activation function, which prevented negative runoff 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 runoff and precipitation data, while z-score standardization was used for the temperature. The ~~model was~~ 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, 2017)~~that~~, which minimized a combined objective function consisting of the KGE and NSE metrics ~~. For a detailed explanation of the employed loss function, refer to Appendix ??~~(see Appendix C).

The architecture presented in Fig. 4 was used to develop two model variants. The first variant, ~~HLSTM-PBHM, included~~ eLSTM, solely included the observed runoff in the hindcast as well as the simulated runoff in both the hindcast and forecast

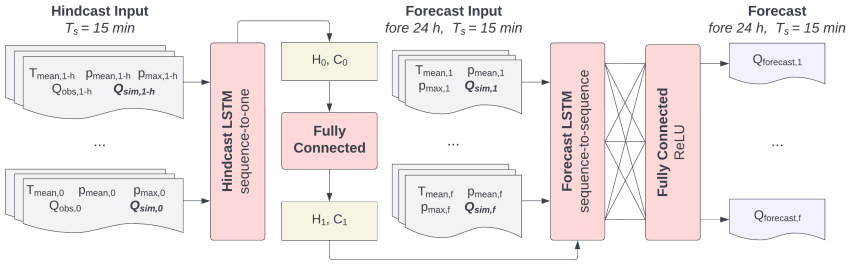


Figure 4. LSTM architecture. The optimized LSTM models incorporated the ~~meteorological quantities p_{mean}, p_{max}, and t_{mean}~~ PBHM's simulations Q_{sim} in both the hindcast and forecast periods ~~. Furthermore,~~ as well as the observed runoff ~~Q_{obs} was used as a feature for~~ at the ~~hindcast LSTM~~gauging station Q_{obs}. ~~HLSTM-PBHM~~ The PBHM-HLSTM exclusively incorporated ~~results from~~ the ~~PBHM Q_{sim}~~ 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). Noteworthy, h and f refer to the hindcast and forecast periods, respectively.

periods. The second model variant, PBHM-HLSTM, additionally included meteorologic forcings. Specifically, the ~~meteorological forcings given in Table A1. These forcings included the~~ catchment's mean and maximum precipitation ~~and~~ as well as its mean temperature in both the hindcast and forecast periods were used. Additionally, ~~HLSTM-PBHM~~ PBHM-HLSTM incorporated runoff observations in the hindcast period and the PBHM's results in both the hindcast and forecast periods, respectively. ~~The second model variant, HLSTM, included similar features as HLSTM-PBHM. However, for this model variant, the results of the PBHM were not included.~~

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 Gradient Method (IG, Sundararajan et al., 2017). Noteworthy, 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:

$$IG_i^{approx}(\boldsymbol{x}) = (x_i - x_i') \times \sum_{k=1}^{m} \frac{\partial F\left(\boldsymbol{x'} + \frac{k}{m}(\boldsymbol{x} - \boldsymbol{x'})\right)}{\partial x_i} \times \frac{1}{m} \tag{1}$$

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 i^{th} dimension, i.e., at the i^{th} 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, for evaluating the performances 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 five previously unseen testing years. 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-hour forecast horizon as well as the individual values corresponding to the 96 forecast steps. Besides that, we also monitored the propagation of the mean absolute error (MAE) and the variability of 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 at flood event runoff? 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 runoff in both timing and magnitude. For this purpose, we computed the median peak magnitude error as well as the median temporal offset across all forecasts in a predefined evaluation window. To add to this, we also evaluated the distribution of the MAE and the variability of the absolute errors. This was done analogously to the methodology presented in the previous point, but for the highest 2 % of the runoff 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 at flood event runoff, 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 at flood event runoff 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 runoff peak. For instance, how important is the observed runoff when the maximum peak is three steps away from the forecast origin, t_0.

4 Results

4.1 ~~Annual model performance~~

4.0.1 ~~Average performance and generalization capability~~

4.1 Overall model performance

4.1.1 Annual average model performance

Evaluating the average annual model performances for the NSE, KGE, PBIAS, and FHV metrics showed that all ~~investigated model variants were able to enhance~~ model variants improved upon the underperforming PBHM's results. Each model's annual ~~NSE, KGE, and PBIAS~~ performance metrics, averaged over the 24-hour forecast horizon, are reported in Table ~~??~~1.

The results revealed that the LSTM-based models excelled in terms of NSE and KGE, which was found to be especially true for the PBHM-HLSTM. For instance, ~~they were able to elevate the average NSE values of~~ the ~~PBHM from 0.19 to at least 0.87~~ PBHM-HLSTM was able to achieve an average NSE value of 0.92 in ~~2013.~~ 2013 compared to the 0.19 of the original PBHM. Even in the worst-performing year, 2017, the ~~LSTM-based models were~~ PBHM-HLSTM was able to elevate the average KGE and NSE values of the PBHM from 0.19 and -4.24 to ~~well above 0.87 and 0.74. Contrary to that, the forecasts obtained by ARIMA displayed a particularly low PBIAS error~~ 0.83 and 0.70, respectively. Overall, the PBHM-HLSTM was found to outperform both ARIMA and the eLSTM in terms of 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's performance was found to be outstanding when compared to the other model variants. We found that this ~~can be attributed to the fact that our ARIMAmodel performed exceptionally well for forecasts that~~ was due to ARIMA's exceptional performance in cases where the forecasts followed a clear ~~trend or pattern . Noteworthy,~~ pattern or trend, which in hydrologic model applications is often the case in ~~hydrologic modeling applications , this is the case for most forecasts throughout the year, i.e., in~~ baseflow conditions. ~~In these instances, our ARIMA model produced near-perfect forecasts, reflected in the close-to-zero PBIAS values.~~ 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 ~~of~~ in the forecasts obtained by ARIMA. The most straightforward way to identify these shortcomings was by dissecting

Table 1. Average annual model performance comparison. ~~Included~~ Shown are the ~~annual averages of the~~ efficiency metrics KGE, NSE, ~~and~~ PBIAS, and FHV. All metrics are averaged ~~across~~ over the entire forecast horizon and are reported for ~~all years used for evaluation~~each testing year. The best values per metric and year are highlighted in **bold**.

~~To assess the forecast models' generalization capabilities, we computed the absolute differences (Δ) between the hydrologic metrics obtained in the validation and testing~~

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 runoffs. As expected, the KGE's bias term for the ARIMA forecasts was close to perfect. Also, the variability term did not signal systematic shortcomings compared to the ~~LSTM results~~ results of the LSTMs. 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 ~~Performance~~ Average model performance over lead time

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

As anticipated, both ~~the ARIMAand LSTM models~~ ARIMA's and the LSTMs forecasts surpassed the PBHM's results across ~~all~~ most evaluated metrics and years. ARIMA, in particular, demonstrated an exceptional performance ~~in terms of PBIAS .~~ for both bias metrics, PBIAS and FHV. The only exception was found to be ARIMA's high FHV in 2015. Also in terms of NSE and KGE, ARIMA showed ~~an~~ outstanding forecast accuracy for the first couple of forecast steps. However, this initial accuracy showed to decline quickly with increasing lead time. This ~~fact~~ became particularly evident in 2017 when ARIMA's initial KGE dropped from 0.98 in the first prediction step to ~~0.52~~ 0.60 in the last. An even more significant performance decrease was observed for the NSE metric, ~~where ARIMA achieved a~~ for which ARIMA achieved an initial value of 0.97 in the first step but

merely ~0.10 0.29 in the last. Compared to the forecasts ARIMA, the LSTM models displayed a different forecast behavior. First, the bias metrics of both LSTM models were mostly higher when compared to those obtained by ARIMA, the LSTM models generally displayed a lower accuracy in the 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 LSTMs consistently performed worse than ARIMA in the first forecast steps, as suggested by both the KGE and NSE metrics. However, they were able contrary to ARIMA, they tended to mostly sustain their initial accuracy across the entire forecast horizon. Even This was found to be most pronounced for the NSE, though it was also observed to a lesser degree for the KGE. For instance, even in the worst-performing year, 2017, the LSTM models were PBHM-HLSTM was able to uphold at least a KGE of approximately 0.82 and an NSE of 0.64.

The results presented in Fig. 5 demonstrate the superiority of the LSTM-based models in obtaining accurate forecasts for longer lead times 0.62 and a KGE of 0.73 across the 24-hour 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. In terms of NSE, the LSTMs outperformed ARIMA after a maximum of 16 lead time steps (4.00 hours) and 5 lead time steps (1.25 hours) on average. For the KGE, the required time for the LSTMs to surpass ARIMA was generally higher. This outcome was no surprise given that the KGE metric includes a direct measure for the bias, for which ARIMA demonstrated near-perfect prediction accuracy. The reason for that was already explained in Sect. **??**.

The comparison of the results of both LSTM variants, HLSTM-PBHM and HLSTM, did not reveal substantial advantages of one over the other. Both models outperformed each other in certain years and evaluation metrics. This implies that the presented results do not offer clear evidence of whether the LSTM model benefited from the inclusion of the PBHM's results or notBesides a few exceptions where both models performed on par, the PBHM-HLSTM outperformed the eLSTM across all years used for evaluation in this regard. This clearly highlights the added benefit of adding meteorologic forcings into model development.

4.1.3 ARIMA and HLSTM-PBHM comparison

To allow for a direct comparison between both correction models, ARIMA and HLSTM-PBHM, we evaluated the number of superior forecasts obtained by both model variants, quantified by the normalized win ratio. For this evaluation we considered the model variant with the lower In addition to the presented metrics used for model evaluation, we also measured the forecast performance by means of the mean absolute error (AE) to be superior to the other. Additionally, we evaluated the stability of the forecasts obtained by both model variants. The forecast stability was gauged by monitoring the development of the AE and its variability MAE) and the standard deviation of the absolute errors ($\sigma(AE)$). Both measures were evaluated across the forecast horizon . The variability of the forecasts was quantified by means of the standard deviation. The results of this evaluation are and are shown in Fig. 6. Analogously to the results presented in Fig. **??**.

When solely considering the normalized win ratio, ARIMA showed to outperform HLSTM-PBHM more often than not.Especially, 5, the trend of the mean absolute error (Fig. 6, left) displays ARIMA's high forecast accuracy in the first forecast step

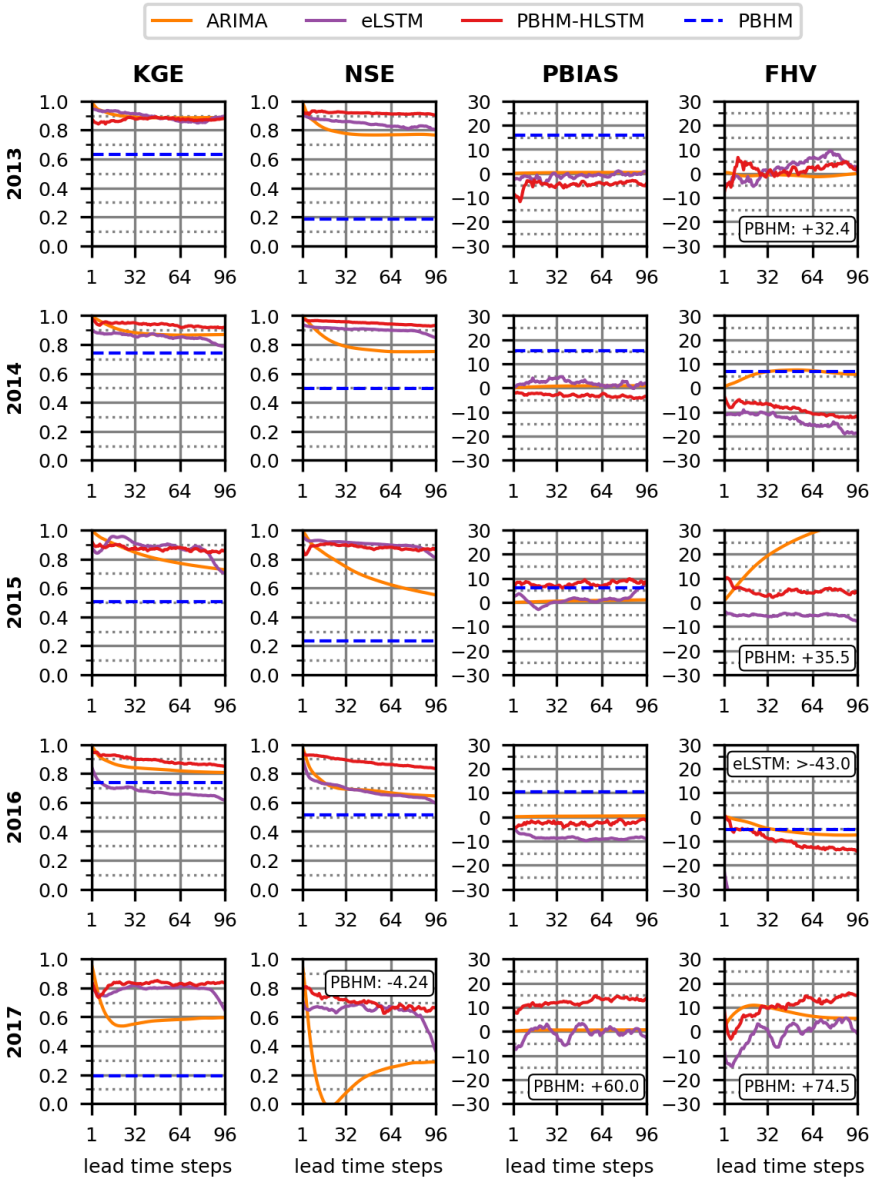


Figure 5. Development of the KGE, NSE, ~~and~~ PBIAS and FHV metrics over the 24-hour (96 lead time steps) forecast horizon. ~~The evaluations include~~ Included are all developed model variants and all testing years.

~~ARIMA yielded better results in a minimum of 75 % of the total forecasts in 2014 and up to 91 % in 2016. Interestingly, even for longer lead times, ARIMA outperformed the LSTM in many instances throughout the year. Again, this can be attributed to the fact that ARIMA's performance was near-perfect for forecasts that followed a clear trend or pattern (see~~

Sect. ??). Investigating the development of the mean absolute error suggested a widely equal performance of the ARIMA and HLSTM-PBHM forecasts.Exceptions for this were found to be the years 2014 and 2016, where HLSTM-PBHM achieved a more favourable mean error. Contrary to that, the forecasts obtained by ARIMA displayed a much higher standard deviation, especially for longer lead timescouple of prediction steps. However, it also reaffirmed its gradual decline in accuracy. Contrary to that, 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 of the errors shown in Fig. 6 (right). Unlike the LSTM-based models, the forecasts generated by ARIMA demonstrated significant variability in their errors. This indicates that ARIMA in certain instances produced considerably worse forecasts compared to the LSTM. In this regard, HLSTM-PBHM surpassed ARIMA while ARIMA produced highly accurate forecasts in some cases, it often yielded predictions that deviated substantially from the actual outcomes, especially for further ahead forecasts. In contrast, both LSTM variants achieved a considerably lower error variance. Quantitatively, the LSTM-based models provided more reliable forecasts on average after four three forecast steps (1.00 hoursi.e., 45 minutes).

4.2 Performance for elevated river runoff

4.2.1 Peak timing and magnitude errors

For assessing the performances of our forecast models at flood event runoff, we determined the models' median peak magnitude and timing errors for the two largest flood events in each year. The magnitude error, e_{peak}, quantifies the median offset between the maximum observed and simulated peak runoff across the evaluation window in percent. Similarly, the timing error Δt measures the median temporal offset between the maximum observed and simulated peak runoff in number of time steps. Positive magnitude errors indicate model overestimation, while negative values suggest an underestimation. As for the timing errors, negative values indicate that the model predicted the maximum peak runoff earlier than observed, and positive values indicate the opposite. The results of this evaluation are presented in Table 2.

Upon initial inspection, the presented results highlight the evaluation of the peak magnitude and timing errors reaffirmed the deficiencies of the PBHM in capturing both the peak magnitude and its timingthe flood runoff dynamics. Especially, the substantial timing errors suggest shortcomings of the model in adequately depicting the characteristics of the flood event hydrographs. In terms of magnitude error, the PBHM exhibited a median magnitude error of -44.7achieved a median value of -49.9 %, predominately underestimating the observed peak runoff. Interestingly, both ARIMA and the LSTMs showed only modest improvements compared to the PBHM, with median errors of -38.5 %, -28.6 %, and -22.2 % for ARIMA, HLSTM-PBHM, and HLSTM, respectively. Considering this, ARIMA was able to elevate the PBHM's Arguably, none of the investigated model variants was 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 by approximately 6 %, HLSTM-PBHM by 16 %, and HLSTM by 22 %, respectively. Although, HLSTM was able to achieve the highest relative improvement compared to the

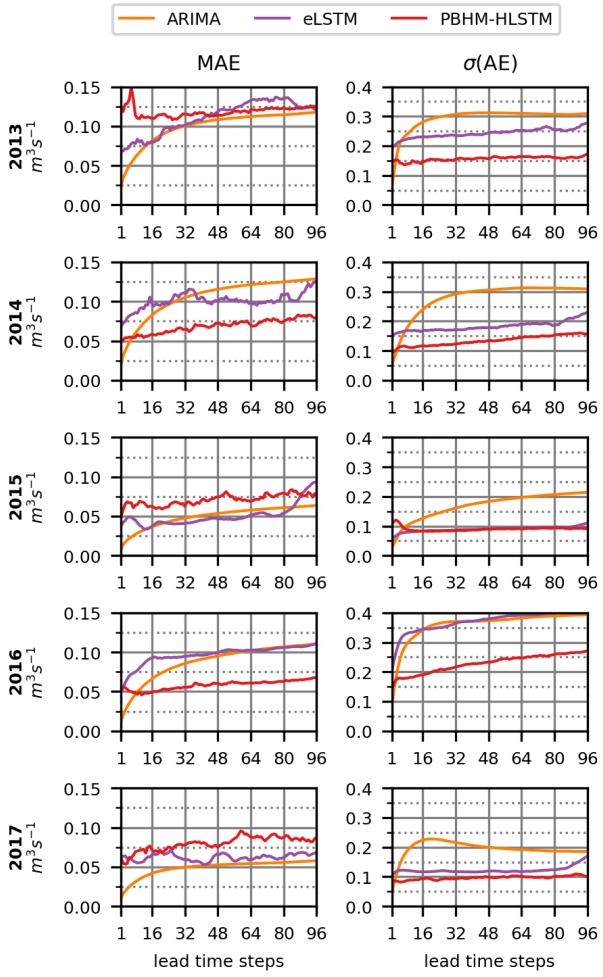


Figure 6. ~~ARIMA and HLSTM-PBHM forecast comparison~~ Development of the absolute errors for all ~~forecasts~~ flows. Shown are the MAE and the standard deviation σ of the AE per testing year for the 24-hour forecast horizon (96 time steps).~~(Left) Normalized win ratio (i.e, ratio of superior model forecasts). (Right) Propagation of the absolute error. Shown are the mean μ and the standard deviation σ of the absolute errors. .~~

~~PBHM, its magnitude errors were still large~~of -27.5 %. Interestingly, the worst performance in this regard was shown by the eLSTM and ARIMA's results lying in between. 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 ~~errors~~error, the ARIMA-corrected forecasts showed no improvement compared to the PBHM's ~~results. In fact, quite the opposite was observed. Whilst the PBHM achieved an absolute median timing error of 27 time steps, a value of 37 was achieved by ARIMA. Contrary to that , both LSTM variants were able to significantly reduce the~~ original forecasts.

Table 2. Comparison of the median peak magnitude error e_{peak} (in percent) and timing ~~errors~~ error Δt (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 runoff	PBHM		ARIMA		eLSTM		PBHM-HLSTM	
		$(m^3 s^{-1})$	e_{peak} (%)	Δt	e_{peak} (%)	Δt	e_{peak} (%)	Δt	e_{peak} (%)	Δt
2013	1st	15.00	-90.3	31	~~-75.3~~-71.7	47	~~-44.7~~-78.0	~~29~~59	**~~-16.2~~-38.4**	**~~0~~2**
	2nd	10.02	~~+13.3~~**+13.3**	20	**~~-3.5~~**-26.1	19	~~-49.8~~-40.0	~~8~~**-1**	~~+17.4~~-27.5	~~3~~-5
2014	1st	7.27	~~+3.5~~ **+3.5**	18	**~~+2.1~~**+4.5	~~18~~19	~~-33.5~~-27.0	**~~-1~~0**	~~-46.1~~-27.8	~~7~~-2
	2nd	6.23	+23.3	16	~~+21.3~~**+19.1**	16	**~~-20.0~~**-24.8	**~~-2~~2**	~~-37.6~~-26.5	~~2~~4
2015	1st	5.85	-62.5	36	~~-46.9~~-29.9	36	~~-22.5~~-66.7	**~~0~~**10	**~~-15.9~~-25.2**	~~-2~~2
	2nd	3.33	**+4.7**	49	+~~6.2~~17.4	~~48~~44	~~-9.8~~-32.7	**~~2~~0**	~~-18.9~~-13.4	~~2~~3
2016	1st	17.94	-73.6	26	~~-47.4~~**-45.9**	~~47~~29	~~-23.8~~-77.1	~~2~~5	**~~+23.4~~**-68.8	**~~1~~3**
	2nd	9.99	~~-39.4~~-45.4	~~-96~~**18**	~~-33.9~~-56.7	~~-89~~20	~~-67.5~~-65.5	~~4~~27	**~~-25.4~~-43.9**	~~2~~68
2017	1st	9.21	-49.9	25	**~~-43.1~~**-48.9	~~37~~48	~~-56.4~~-54.5	~~63~~3	~~-62.0~~**-17.3**	**~~3~~-1**
	2nd	7.37	-63.1	28	~~-44.7~~-32.6	~~29~~31	**~~-16.2~~**-59.9	**~~0~~**6	~~-47.6~~ **+8.4**	~~57~~**1**
all folds	1st		-62.5	26	-45.9	36	-66.7	5	**-27.8**	**2**
	2nd		**+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**

In contrast, the eLSTM was able to majorly reduce the timing ~~errors in the forecasts. More specifically, both LSTM variants achieved a median absolute~~ errors in the forecasts. Considering the fact that both models were supplied with the same input data 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 ~~, equivalent~~ across all events, which corresponds to 30 minutes ~~.~~ in this study. In contrast, the median timing error of the eLSTM was found to be 5 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 ~~ARIMA and HLSTM-PBHM comparison~~ Performance over lead time for elevated ~~river~~ runoff

~~Similar~~ Analogously to the results presented in ~~Sect. ??,~~ Fig. ~~??~~ ~~shows the normalized win ratio~~ 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, but only evaluated for the largest ~~5~~2 % of the ~~annual runoff. Furthermore, the propagation of the absolute error and its variability are shown~~runoff. The results of this investigation are shown in Fig. 7.

~~The presented results highlight the superiority of HLSTM-PBHM in improving the forecast accuracy at elevated river runoff, particularly for longer lead times. This manifested in both the normalized win ratio (except for 2015) and also in the propagation~~

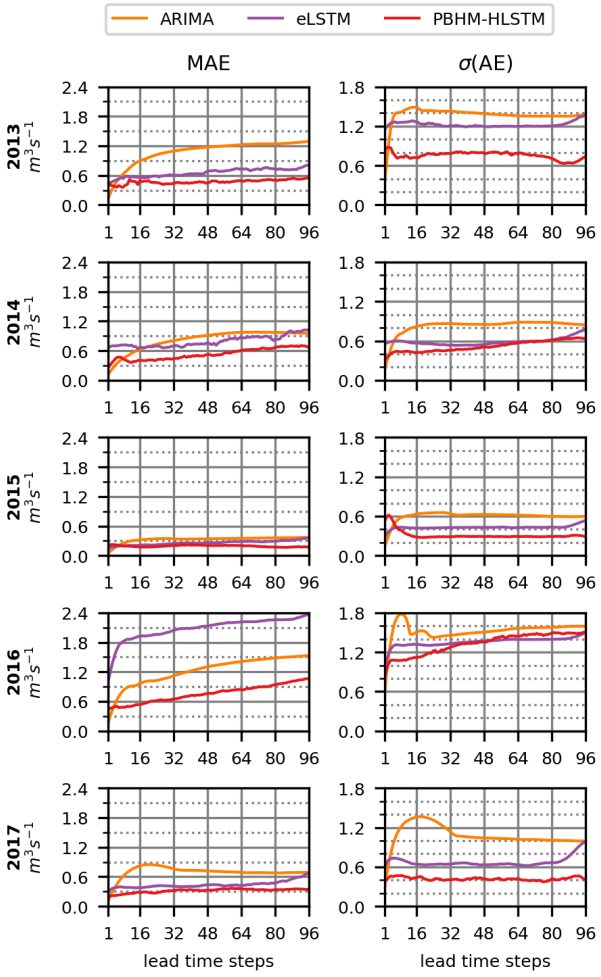


Figure 7. Development of the absolute errors for the largest 2 % of the runoff. Shown are the MAE and the standard deviation σ of the AE per year for the 24-hour forecast horizon (96 time steps).

~~of the mean absolute error and its standard deviation. Except for 2015,~~ While ARIMA often outperformed or matched the MAE of the LSTMs when considering all flows (see Fig. 6, left), the evaluation of the largest runoff values clearly demonstrates the superiority of the LSTM-based models (see Fig. 7, left). Particularly the PBHM-HLSTM, but to a lesser degree also the eLSTM, was able to achieve a considerably lower MAE compared to ARIMA. A similar picture was drawn by the variance of the absolute errors (see Fig. 7, right) for which the PBHM-HLSTM displayed considerably lower values than both ARIMA and the eLSTM.

Table 3. Normalized feature importances 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 or equal to 0.01 are omitted to increase readability.

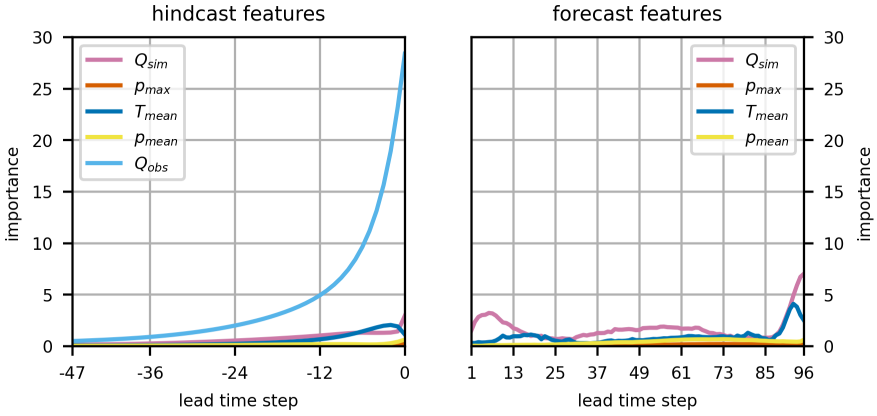
year	hindcast features					forecast features			
	Q_{sim}	p_{max}	T_{mean}	p_{mean}	Q_{obs}	Q_{sim}	p_{max}	T_{mean}	p_{mean}
2013	0.09		0.03		**0.55**	0.19		0.08	0.04
2014	0.08		0.04	0.02	**0.55**	0.18		0.08	0.04
2015	0.08		0.08		**0.51**	0.16		0.12	0.03
2016	0.07		0.07	0.02	**0.49**	0.18		0.10	0.04
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

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 LSTMs is shown in Fig. 8. As anticipated, the ~~LSTM's absolute error for longer lead times was almost half of ARIMA's errors. A similar trend was observed for the standard deviations. Noteworthy, also for elevated runoff, ARIMA's forecastaccuracy was higher in the first time steps compared to HLSTM-PBHM. On average, the LSTM required four forecast steps to surpass the results of ARIMA.~~ PBHM-HLSTM heavily relied on past runoff 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 to that, 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 runoff was found to have the second highest impact on the forecasts, right after the observations. Particularly in the forecast period, the simulated runoff influenced the final predictions considerably.

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



~~ARIMA~~

Figure 8. Importance of all input features summed over all testing years. Shown are the feature importances of the hindcast features (left) and the forecast features (right)

4.3.2 Sensitivity at peak runoff

To investigate the importance of the individual features at flood event runoff, we exclusively evaluated the integrated gradients for the two largest flood events of each year. Figure 9 shows the importances of the hindcast (left) and forecast (right) features for various distances of the predicted runoff 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 runoff. This comes as no surprise as the observed runoff at t_0 should be a reasonable predictor for the runoff 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 runoff still had a rather high impact on the forecast. As for the precipitation features, p_{mean} and ~~HLSTM-PBHM forecast comparison for the largest 5 %~~ 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. Further investigating the mean precipitation feature revealed additional insights. First, the hindcast p_{mean} showed to have a high influence when the peak was close to t_0, which then decayed exponentially with increasing distance of the peak runoff 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 showed 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. Also for predictions at flood event runoff, it showed 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 widely equal to that of the maximum precipitation p_{max} and the mean temperature T_{mean}.

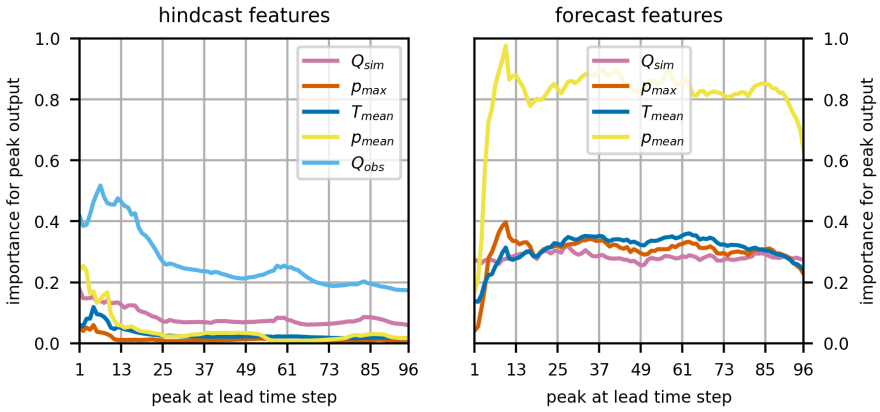


Figure 9. Importance of ~~the annual runoff values~~ input features for the ~~24-hour~~ peak prediction in the forecast ~~horizon (96 time steps). (Left) Normalized win ratio (i.e, ratio of superior model forecasts). (Right) Propagation of~~ window summed over the ~~absolute error~~two largest flood events per year. Shown are the ~~mean μ and the standard deviation σ~~ feature importances of the ~~absolute errors~~hindcast features (left) and the forecast features (right).

Table 4 summarizes the normalized feature importances for the two highest 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 predicting flood event runoff, with the only exception being the observed runoff in the hindcast. As for the Q_{sim}, p_{max}, and T_{mean} features in the forecast, all showed to have a more or less equal influence on the final flood event 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 LSTMs for enhancing the forecast accuracy of PBHMs employed in operational flood forecasting systems. ~~For this purpose~~Following the approaches of Gauch et al. (2021) and Nevo et al. (2022), we developed ~~an LSTM model (HLSTM-PBHM) that was largely inspired by the~~ our LSTM models using a hindcast-forecast architecture~~presented by Nevo et al. (2022). This specific architecture was selected~~ . This architecture was chosen as it facilitates an effective 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 ~~for distinguishing meteorological~~ the model to distinguish between meteorologic forecasts and analyses, potentially enabling ~~the model~~ it to learn from their differences. Furthermore, ~~with~~ the here presented cross-validation strategy ~~, we established a framework for a~~ enables a seamless continuous improvement of the model as new ~~observational data become~~ data becomes available. To ~~showcase the proposed model's effectiveness, we rigorously compared its forecasting capabilities to those of a~~

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 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	1st	0.02		0.02		0.09	0.04	0.11	0.11	**0.59**
	2nd	0.12		0.02	0.04	0.20	**0.28**	0.02	0.12	0.19
2014	1st	0.06		0.02	0.02	**0.32**	0.22	0.02	0.11	0.23
	2nd	0.07		0.02	0.06	0.20	**0.24**	0.06	0.15	0.21
2015	1st			0.03	0.03	0.05	0.05	0.12	0.18	**0.52**
	2nd	0.02				0.11	0.16	0.11	0.22	**0.37**
2016	1st	0.02				0.14	0.10	0.18	0.15	**0.36**
	2nd	0.12			0.03	0.21	0.24	0.02	0.10	**0.26**
2017	1st					0.02	0.07	0.27	0.16	**0.44**
	2nd						0.03	0.24	0.15	**0.54**
all folds	1st	0.02				0.12	0.10	0.17	0.14	**0.41**
	2nd	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**

~~more~~ assess the benefits of the LSTM-based forecasts, we developed two LSTM model variants and compared their forecast skills to that of a conventional ARIMA model, using one underperforming PBHM as a case study. ~~Particularly interesting was how the proposed LSTM model (HLSTM-PBHM) improved the forecast accuracy~~To ensure comparability between the LSTM and ARIMA approaches, one LSTM (eLSTM) was restricted to use the same data as ARIMA, while the other incorporated additional meteorologic variables (PBHM-HLSTM). Of particular interest was how the LSTM approach improved prediction accuracy, especially at flood event runoff and for longer lead times ~~and flood event runoff,~~ — both being recognized weaknesses of ARIMA ~~models~~for cases where the underlying PBHM provides poor initial estimates.

When comparing the forecasts obtained by ~~both correction models, ARIMA and HLSTM-PBHM,~~ the LSTM and ARIMA models, we observed that both ~~had their~~ methods had certain advantages and disadvantages. ARIMA generally ~~showed~~ demonstrated a very high accuracy in the first forecast steps. However, this initial accuracy often showed to decline quickly 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 model typically~~ LSTMs generally exhibited a larger error in the first steps but ~~it was~~ were able to mostly sustain ~~its initial accuracy across the~~ their initial accuracy over the 24-hour forecast horizon. ~~We also observed that for longer lead times the LSTM yielded much more reliable forecasts. In most instances, it achieved a lower absolute error and also displayed a lower variability in these errors. This was particularly evident in elevated runoff conditions~~This became

particularly evident when observing the variance of the absolute errors. Both LSTM models, particularly the PBHM-HLSTM, displayed a considerably lower error variance compared to the results obtained by ARIMA. This suggests that they produced exceptionally poor forecasts less often. Interestingly, ARIMA performed exceptionally well in terms of PBIAS. The reason for that was found in ARIMA's ~~exceptionally~~ high accuracy for forecasts that followed a clear trend or pattern, which ~~appears~~ in hydrologic model applications occurs most often during ~~base flow~~ baseflow conditions.

When ~~considering~~ focusing solely on the forecast skills at flood event runoff, the ~~LSTM~~ LSTMs clearly outperformed ARIMA. This ~~was indicated by both the KGE' s Gupta et al. (2009) correlation term and also the obtained errorsat selected flood events. Although both the LSTM and ARIMA models were not~~ became particularly evident when investigating the models' timing errors, i.e., the temporal offset between the maximum observed and simulated peak runoff. While both LSTM variants were able to significantly ~~improve the PBHM's magnitude errors, the LSTM was able to significantly reduce its timing errors . Contrary to that, ARIMA's timing errors were even larger than those of the original PBHM~~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 ~~reported~~ shown by Liu et al. (2015). ~~Overall we found that our LSTM model outperformed ARIMA in all aspects we consider relevant for operational flood forecasting~~As for the magnitude errors, i.e., ~~a more accurate representation of flood event runoffand more reliable forecasts for longer lead times.~~ the difference between the maximum observed and simulated runoff, 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 runoff 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.

~~Reflecting on the~~ Considering the comparably high performance of the PBHM-HLSTM in this study and more generally the remarkable capabilities of LSTM models in predicting river runoff (e.g., Kratzert et al., 2019b)~~inevitably prompts the question~~ , it raises the question 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 ~~whether these advancements render PBHMs obsolete in operational flood forecasting. In a study, Frame et al. (2021) demonstrated that, in many instances, a standalone LSTM outperformed two hybrid LSTMs that included the PBHMs results . In light of these findings, we critically scrutinized our approach by comparing its forecast skills to those of an LSTM variant (HLSTM) that did not include~~ the PBHM, particularly its forecasts. In fact, the PBHM's ~~results. This investigation was conducted to test whether the LSTM can fully replace the underperforming PBHM. For the here presented test case, we found no distinct evidence of whether our LSTM benefited from including~~ runoff predictions were identified as the second most important feature, following the observed runoff. While the PBHM-HLSTM at flood event runoff did to some extent also rely on the PBHM's forecasts~~or not. Both model variants yielded viable results occasionally outperforming each other in some of the years and metrics used for evaluation. The widely equal performance of both model variants suggests that the decision of which strategy should be employed has to be made under careful consideration of the forecasting system's requirements~~, 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 at flood event runoff.

~~Nevertheless, it is crucial to consider certain aspects when implementing solely data-driven models into operational forecasting systems: (I) Training on erroneous data: In the scenario presented, there are two primary sources of uncertainty. Firstly, the training data may carry systematic uncertainties, such as an underestimation of the rainfall intensity by the meteorological model. Secondly, there is the possibility of erroneous gauging data (target data), which can for example result from translating the measured river stage to runoff (e.g., ?). In instances of erroneous training data, data-driven models might be adept at learning any systematic errors embedded in the data, presenting a viable alternative to PBHMs. Conversely, in the case of erroneous gauging data, data-driven models may still yield seemingly usable results, having learned from these errors, while PBHMs struggle to adapt and may signal potential issues. (II) Out-of-sample predictions: In this study, we have demonstrated that our data-driven models were able to achieve a higher generalization capability compared to the underperforming PBHM. However, this might not be true in instances where the underlying catchment processes are captured well by the PBHM, particularly if limited data is available for training the LSTMs (e.g., ?). (III) Limited availability of system states : The information of system states in the catchment is limited when employing solely data-driven models~~When employing forecast-enhancing models in operational flood forecasting systems, several important considerations must be taken into account. First and foremost, such models are no all-in-one device suitable for every purpose. Although the here presented PBHM-HLSTM 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, ~~many operational forecasting systems rely on information of the~~ often these system states, e.g., the state of the snow cover, the soil moisture, or spatially distributed information of the runoff~~in the catchment. Often these states~~ , function as an additional decision criterion for the system's operator ~~or are required for the implementation of~~ 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 parametrization, 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 runoff). Notably, in contrast to PBHM's, data-driven models (e.g., LSTMs) 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 ~~here presented LSTM models~~ PBHM-HLSTM model presented here already achieved a comparably high forecast accuracy, there ~~a potential exists~~ exists potential for future enhancements. ~~Firstly,~~ First, refining the pre-processing phase~~can be intensified. Specifically, a more careful approach to~~ , especially through more targeted feature engineering could ~~increase~~ further enhance the model's ~~quality by providing more relevant and informative features included in training. Secondly~~predictive capabilities. Second, the target data (gauge runoff) can be diagnosed. For instance, adopting the probe technique presented by Lees et al. (2022) could be used 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 ~~leveraging the individual strengths of each model~~further increasing the model's prediction accuracy, particularly in the first forecast steps.

6 Conclusions

In this study, we ~~proposed a forecast correction method, based on a hindcast-forecast LSTM network (HLSTM-PBHM). The efficacy of this proposed method was demonstrated by comparing its forecast accuracy to results obtained by a conventional ARIMA model, utilizing one underperforming PBHM as a case study. Additionally, we compared both correction strategies to a standalone LSTM that did not incorporate the PBHM~~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' performances for extended lead times and particularly at flood event runoff, 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 runoff and the runoff generated by the PBHM, while the other, PBHM-HLSTM, additionally incorporated meteorologic variables. Furthermore, we assessed the added value of the underperforming PBHM's results ~~(HLSTM)~~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-hour forecast horizon.
- ARIMA showed shortcomings in forecasting flood event runoff. Specifically, it failed to accurately predict the timing and the maximum peak runoff 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 runoff. 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~~, particularly if the underlying PBHMis underperforming~~.

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}).

			year						
parameter	**statistic**	**unit**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**
Q_{obs}	μ	m^3s^{-1}	0.57	1.01	1.21	1.17	0.71	0.83	0.57
	σ	m^3s^{-1}	0.25	0.84	0.68	0.67	0.33	0.68	0.23
	max	m^3s^{-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^3s^{-1}	0.62	1.10	1.40	1.35	0.76	0.92	0.91
	σ	m^3s^{-1}	0.33	0.83	1.02	0.72	0.52	0.70	0.48
	max	m^3s^{-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

Appendix A: Statistics of the input data

A1 Model input data

Table A1 shows the key statistics of the observed and simulated runoff 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 close the residuals follow a Gaussian distribution. For this purpose the Filliben r correlation value (Filliben, 1975) was computed, for which a value close to one 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 for a 5 % significance as 3.84 based on the Chi-distribution given one 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 of heteroscedasticity, even below the critical value for homoscedasticity. Fig. A1 shows a QQ-plot including the Filliben r test statistics for the original and Box-Cox transformed residuals (left), alongside a

scatterplot of the PBHM's residuals against the observed runoff, which includes the test statistics of the Breusch-Pagan test (right).

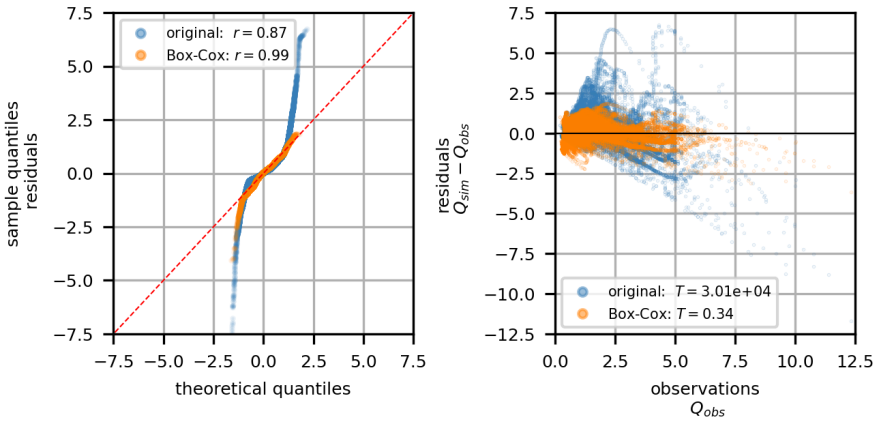


Figure A1. QQ-plot with Filliben r test statistics for the original and Box-Cox transformed residuals (left) Scatterplot of the PBHM's original and transformed residuals against the observed runoff including the test statistics of the Breusch-Pagan test (right).

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 averaged. Fig. A2 includes the original PBHM residuals, the Box-Cox transformed residuals as well as the residuals following one differentiation operation.

Appendix B: Evaluation metrics

B1 Nash-Sutcliffe ~~efficiency~~ Efficiency (NSE)

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

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

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

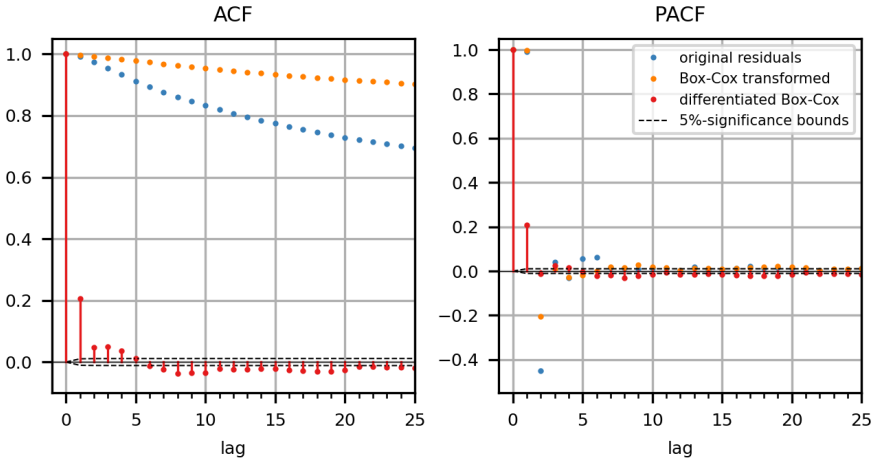


Figure A2. Autocorrelation Function (ACF, left) and Partial Autocorrelation Function (PACF, right) for the original PBHM model residuals, the Box-Cox transformed residuals as well as the residuals following one differentiation operation.

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 of the flow. In this study, we utilized the modified Kling-Gupta Efficiency (Kling et al., 2012), which can be written as:

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

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

B3 Percent ~~Bias~~ bias (PBIAS)

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

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

where ~~$Q_{Obs,t}$ and $Q_{Sim,t}$~~ $Q_{sim,t}$ and $Q_{obs,t}$ is the observed and predicted runoff, 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 zero indicates a widely unbiased model.

Appendix C: ~~Loss function~~

B1 High-segment volume percent bias (FHV)

The FHV quantifies 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:

$$FHV = \frac{\sum_{i=1}^{H}(Q_{sim,i} - Q_{obs,i})}{\sum_{i=1}^{H}(Q_{obs,i})} \cdot 100 \tag{B1}$$

where $Q_{obs,i}$ and $Q_{sim,i}$ is the observed and predicted runoff, 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

For tuning the hyperparameters, we selected a combined objective function f_{obj} consisting of the NSE and KGE metrics. ~~A similar approach was presented by Nevo et al. (2022).~~ The objective function was computed as follows:

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

~~A similar combination of these metrics was also employed as the loss function used in training the models. To mitigate potential issues arising from the unbounded lower limit of the NSE and KGE, both metrics were normalized such that their values fall between zero and one, as also suggested by Nossent and Bauwens (2012). To be compatible with the minimization approach of the chosen optimizer, the values were also inverted, meaning that zero indicates~~ where zero would indicate a perfect fit by the model. ~~The resulting loss function can be written as:~~

$$Loss = 2 - normKGE - normNSE$$

Appendix D: ~~Auxiliary information for LSTM hyperparameter tuning~~

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

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

~~Hyperparameter~~Parameter	Search space		~~HLSTM-~~eLSTM* ~~PBHM~~	~~HLSTM~~eLSTM-H48	BPHM-HLSTM-H96	BPHM-HLSTM*
	Min.	Max.				
ID best trial			41	40	48	22
Objective			**0.232**	0.239	0.169	**0.162**
N. of LSTM units	~~2~~44	~~96~~32	~~96~~17	~~46Batch size*~~22	20	~~40004000~~23
Initial learning rate	1e-3	1e-2	~~0.0010~~0.00774	~~0.00223~~0.0090	0.0065	0.00912
~~Retrain epochs*~~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
~~Dropout rate~~Hindcast length	~~0.01~~	~~0.5~~	~~0.370~~96	~~0.418~~48	96	48

*selected for further processing based on tuner objective

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

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 runoff. A summary of these findings can also be found in Table 2.

Code and data availability. The Python code and processed data presented in this study are stored on **Zenodo** (Gegenleithner et al., 2024b). The published data was derived from the following datasets (I) Gauge runoff: Styrian Government, Department 14 – Water Management, Resources and Sustainability (Hydrographic Service of Styria). The data was validated by the provider. The time stamps were converted from GMT+1 to UTC by the authors. (II) Meteorologic data: The meteorologic data was provided by GeoSphere Austria. More specifically, 1x1 km rasters were provided from which we extracted catchment averaged values. Those averaged values are included in the dataset. (III) Hydrologic modeling results: The hydrologic modeling results were obtained from Gegenleithner et al. (2024a). The developed Python code is also available on **GitHub**.

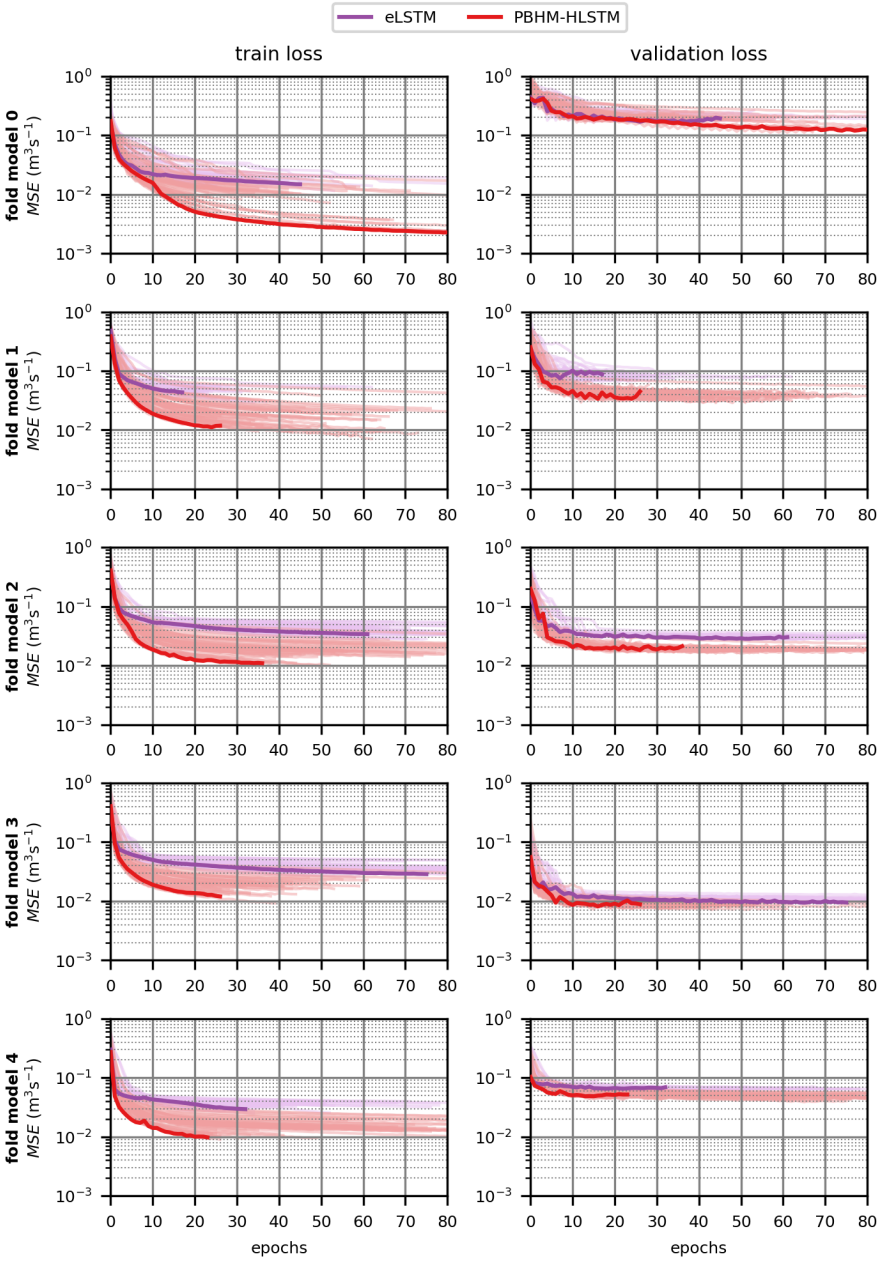


Figure C1. Best models' losses during training and validation.

Author contributions. Sebastian Gegenleithner: Conceptualization, Methodology, Data curation, Writing - original draft preparation. Manuel Pirker: Conceptualization, Methodology, Data curation, Writing - original draft preparation. Clemens Dorfmann: Funding acquisition, Writing - review & editing. Roman Kern: Writing - review & editing. Josef Schneider: Supervision, Writing - review & editing

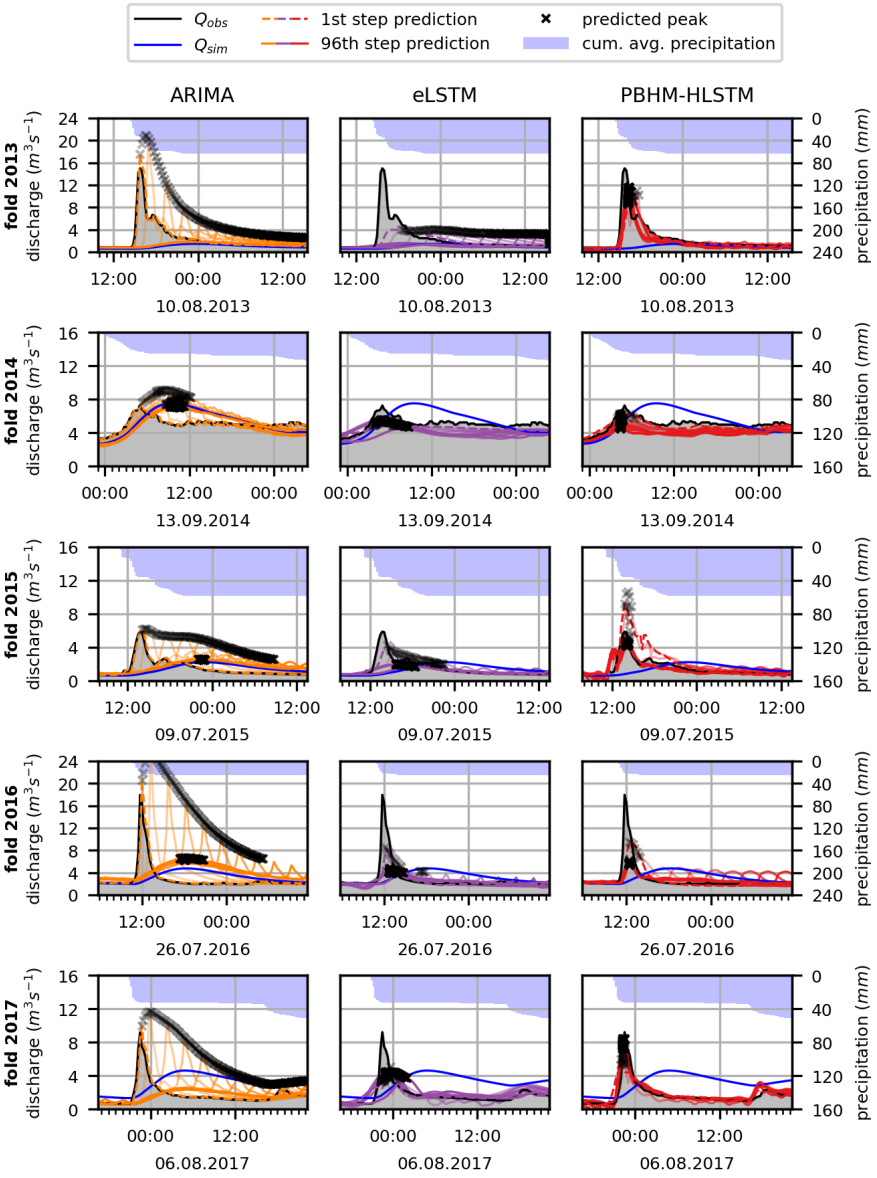


Figure D1. Forecast comparison for the largest runoff event per year. Given are the results of the ARIMA, eLSTM, and PBHM-HLSTM models.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We express our 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.

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.

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, https://www.tensorflow.org/, software available from tensorflow.org, 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, Hydrology and Earth System Sciences, 15, 2327–2347, 2011.

Bergmeir, C. and Benítez, J.: On the use of cross-validation for time series predictor evaluation, Information 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, 2021.

Box, G. and Cox, D.: JR Stat, Soc. Series B, 26, 211, 1964.

Brath, A., Montanari, A., and Toth, E.: Neural networks and non-parametric methods for improving real-time flood forecasting through conceptual hydrological models, Hydrology and Earth System Sciences, 6, 627–639, 2002.

Breusch, T. S. and Pagan, A. R.: A Simple Test for Heteroscedasticity and Random Coefficient Variation, Econometrica, 47, 1287–1294, http://www.jstor.org/stable/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, vol. 2, pp. 963–968, IEEE, 2005.

Chollet, F. et al.: Keras, https://keras.io, 2015.

Engeland, K., Renard, B., Steinsland, I., and Kolberg, S.: Evaluation of statistical models for forecast errors from the HBV model, Journal of Hydrology, 384, 142–155, https://doi.org/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, cited by: 739, 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, JAWRA Journal of the American Water Resources Association, 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, Hydrology and Earth System Sciences, 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, Journal of Hydrology, p. 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, 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, Journal of hydrology, 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 and Forecasting, 26, 166–183, 2011.

Hochreiter, S. and Schmidhuber, J.: Long Short-term Memory, Neural computation, 9, 1735–80, https://doi.org/10.1162/neco.1997.9.8.1735, 1997.

Kalman, R. E.: A New Approach to Linear Filtering and Prediction Problems, Journal of Basic Engineering, 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, 2022.

Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, 2017.

Kling, H., Fuchs, M., and Paulin, M.: Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios, Journal of Hydrology, 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, Environmental Research Letters, 15, 104 022, 2020.

Kratzert, F., Herrnegger, M., Klotz, D., Hochreiter, S., and Klambauer, G.: NeuralHydrology – Interpreting LSTMs in Hydrology, pp. 347–362, ISBN 978-3-030-28953-9, https://doi.org/10.1007/978-3-030-28954-6_19, 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 Resources Research, 55, 11 344–11 354, 2019b.

Land Kärnten: Austria 10m Digital Elevation Model, https://www.data.gv.at/katalog/dataset/land-ktn_digitales-gelandemodell-dgm-osterreich, accessed: 2022-09-22, 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, Hydrology and Earth System Sciences, 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, Journal of Hydrology, 603, 126 888, https://doi.org/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, Natural Hazards, 77, 1161–1182, 2015.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I—A discussion of principles, Journal of hydrology, 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, Journal of Hydrology and Hydromechanics, 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., et al.: Flood forecasting with machine learning models in an operational framework, Hydrology and Earth System Sciences, 26, 4013–4032, 2022.

Nossent, J. and Bauwens, W.: Application of a normalized Nash-Sutcliffe efficiency to improve the accuracy of the Sobol' sensitivity analysis of a hydrological model, in: European Geosciences Union General Assembly 2012, p. 237, 2012.

O'Malley, T., Bursztein, E., Long, J., Chollet, F., Jin, H., Invernizzi, L., et al.: Keras Tuner, https://github.com/keras-team/keras-tuner, 2019.

Rozos, E., Dimitriadis, P., and Bellos, V.: Machine learning in assessing the performance of hydrological models, Hydrology, 9, 5, 2021.

Schellekens, J.: wflow Documentation, 2012.

Seabold, S. and Perktold, J.: statsmodels: Econometric and statistical modeling with python, in: 9th Python in Science Conference, 2010.

Sundararajan, M., Taly, A., and Yan, Q.: Axiomatic Attribution for Deep Networks, in: International Conference on Machine Learning, https://api.semanticscholar.org/CorpusID:16747630, 2017.

Tashman, L. J.: Out-of-sample tests of forecasting accuracy: an analysis and review, International Journal of Forecasting, 16, 437–450, https://doi.org/10.1016/S0169-2070(00)00065-0, the M3- Competition, 2000.

Umweltbundesamt GmbH: Austrian river network, v17, https://www.data.gv.at/katalog/dataset/gesamtgewssernetzfliessgewsserrouten, accessed: 2022-09-22, 2022.

Weerts, A. H. and El Serafy, G. Y.: Particle filtering and ensemble Kalman filtering for state updating with hydrological conceptual rainfall-runoff models, Water resources research, 42, 2006.

Werner, M., Cranston, M., Harrison, T., Whitfield, D., and Schellekens, J.: Recent developments in operational flood forecasting in England, Wales and Scotland, Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling, 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, p. 30, https://www.undrr.org/publication/human-cost-disasters-overview-last-20-years-2000-2019, accessed: March 2024, 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 Resources Research, 44, https://doi.org/https://doi.org/10.1029/2007WR006716, 2008.