https://doi.org/10.5194/egusphere-2024-1030
Preprint. Discussion started: 14 May 2024

Long Short-Term Memory Networks for Real-time Flood Forecast Correction: A Case Study for an Underperforming Hydrologic Model

Sebastian Gegenleithner[1,*], 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, Brockmanngasse 108, 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 events. The majority of these systems rely on process-based hydrologic models (PBHM), which are used to predict future river runoff. To enhance the forecast accuracy of these models, many operational flood forecasting systems implement error correction techniques, which is particularly important if the underlying hydrologic model is underperforming. Especially, AutoRegressive Integrated Moving Average (ARIMA) type models are frequently employed for this purpose. Despite their high popularity, numerous studies have pointed out potential shortcomings of these 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 Long Short-Term Memory (LSTM) network. We showcase the effectiveness of 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 the PBHM's forecasts. Compared to ARIMA, the LSTM achieves a higher forecast accuracy for longer lead times. In terms of flood event runoff, 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 performance of the proposed and standalone LSTM models.

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

https://doi.org/10.5194/egusphere-2024-1030
Preprint. Discussion started: 14 May 2024

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).

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. (2019) 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 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 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 the model's forecasts. Considering the available literature, the most relevant correction 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). (II) Error correction: These methods use observations of one or multiple state variables, mostly river runoff, to correct 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 potential limitations of these models.

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

https://doi.org/10.5194/egusphere-2024-1030
Preprint. Discussion started: 14 May 2024

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, 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, trained on meteorological data as well as the PBHM's output. 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 hindcast-forecast LSTM correction approach and compare its forecast accuracy to that of a more conventional ARIMA model, using one underperforming PBHM as a case study. Specifically, the selected PBHM has 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, 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 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). 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, particularly for longer lead times? (II) How does the LSTM approach improve the forecast quality for flood events? (III) Does the inclusion of the 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

https://doi.org/10.5194/egusphere-2024-1030
Preprint. Discussion started: 14 May 2024

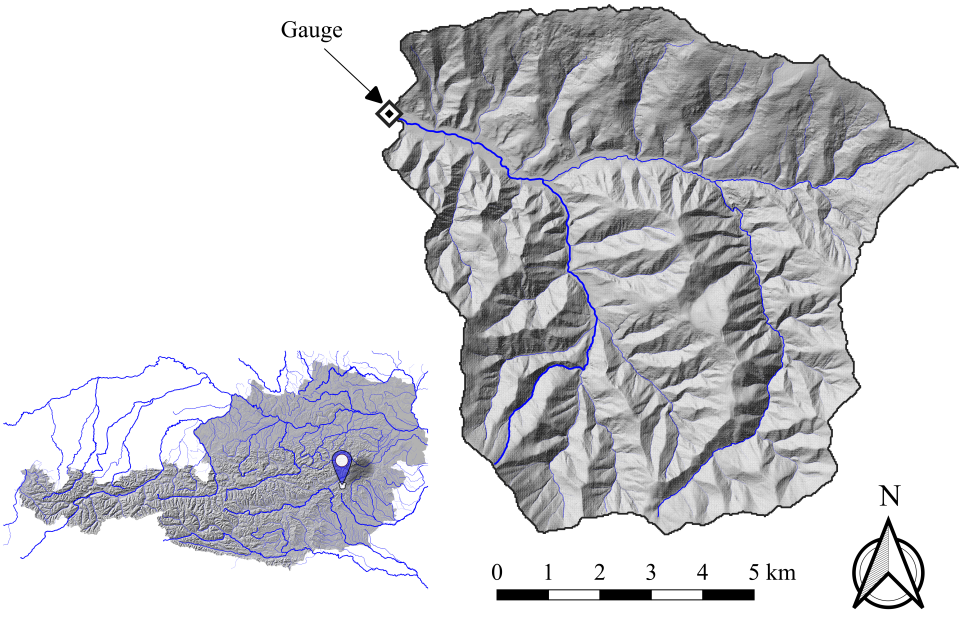

Figure 1. (Bottom left) Location of the study catchment in Austria. (Right) Outline of the study catchment (black line) including the gauging station (black and white diamond) and the main river network (blue lines). This figure was created using the following datasets: Umweltbundesamt GmbH (2022) and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

events occurring at a sub-daily time scale. Fig. 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 was part of a broader study in which various catchments were modeled using a conceptual rainfall-runoff model (Gegenleithner et al., 2024a). This particular catchment was selected for our investigation because the existing rainfall-runoff model failed to provide accurate runoff predictions. For the studied period (2011 - 2017), the existing model merely achieved a Nash-Sutcliffe efficiency (NSE) of 0.43, a Kling-Gupta Efficiency (KGE) of 0.74, and a Percent Bias (PBIAS) of -16.0. For a detailed explanation of these performance metrics, refer to Appendix A. 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, henceforth referred to as Q_{obs}. For the LSTM models exclusively, we also included meteorological 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), were utilized. From the raster data, we extracted the catchment's mean and maximum precipitation, designated as p_{mean} and

https://doi.org/10.5194/egusphere-2024-1030
Preprint. Discussion started: 14 May 2024

Table 1. 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).

			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
	Σ	mm	2159	2877	2777	2847	2198	3222	3076
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

p_{max}, along with its mean temperature t_{mean}. Noteworthy, all processed datasets were available in 15-minute intervals. An overview of the used data and its key statistics is provided in Table 1.

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 PBHM. The first model, ARIMA, relied on forecasting the errors between the simulated and observed runoffs. Subsequently, these errors were used to correct the hydrologic model's forecasts. The second model was based on a hindcast-forecast LSTM network. In contrast to ARIMA, this model directly predicted the runoff by leveraging information on the observed and simulated runoff, along with the meteorological forcings presented in Table 1. Henceforth, we will refer to this model as HLSTM-PBHM. In addition to these models, we developed a variant of HLSTM-PBHM. This variant was implemented 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, HLSTM.

https://doi.org/10.5194/egusphere-2024-1030
Preprint. Discussion started: 14 May 2024

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 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 capabilities.

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.

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' hyperparameters were tuned in the outer loop. Thereby the performance of multiple candidate models was evaluated for the testing sets, and the one that maximized 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 B. A visual representation of the here presented methodology is provided in Fig. 2.

3.1.2 AutoRegressive Integrated Moving Average model

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 ARIMA model presented here 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 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 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:

https://doi.org/10.5194/egusphere-2024-1030
Preprint. Discussion started: 14 May 2024

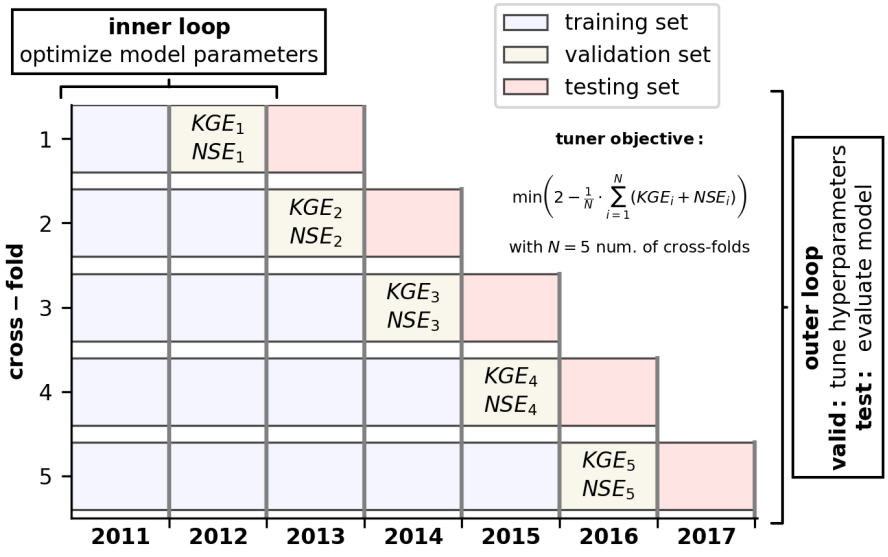

Figure 2. Blocked cross-validation strategy with expanding window setup. The parameters of the models were fitted within the inner loop while the hyperparameters were tuned in the outer loop, utilizing the validation fold of each of the five cross-folds.

$p \in [1, 20]$, $q = p - 1$, and $d \in [1, 2]$. Contrary to other studies (e.g., Broersen and Weerts, 2005), the 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 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 & HLSTM)

Long Short-Term Memory Networks (Hochreiter and Schmidhuber, 1997) are a special form of Recurrent Neural Networks (RNNs). They are specifically designed to address the common issue of vanishing gradients that are often encountered during the training process of RNNs. RNNs process sequential data by maintaining hidden states H that retain information from previous inputs, allowing them to capture temporal dependencies. In addition, 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 Nevo et al. (2022). This architecture involved coupling two distinct LSTM layers, one for the hindcast period and one

https://doi.org/10.5194/egusphere-2024-1030
Preprint. Discussion started: 14 May 2024

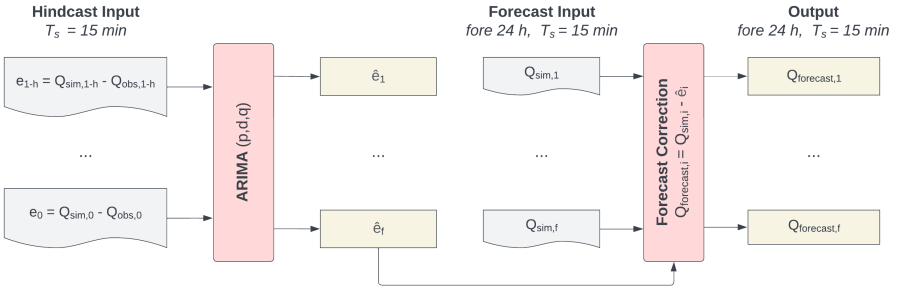

Figure 3. ARIMA architecture. The optimized $ARIMA(p,d,q)$ model utilized the errors between the PBHM's results Q_{sim} and the observed runoff Q_{obs} in the past e to forecast the errors in the forecast period $\hat{e}$. Consecutively, the forecasted errors were used to correct Q_{sim} in the forecast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

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 trained using the Adam optimizer (Kingma and Ba, 2017) that minimized a combined objective function consisting of the KGE and NSE metrics. For a detailed explanation of the employed loss function, refer to Appendix B.

The architecture presented in Fig. 4 was used to develop two model variants. The first variant, HLSTM-PBHM, included the meteorological forcings given in Table 1. These forcings included the catchment's mean and maximum precipitation and its mean temperature in both the hindcast and forecast periods. Additionally, HLSTM-PBHM 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.

https://doi.org/10.5194/egusphere-2024-1030
Preprint. Discussion started: 14 May 2024

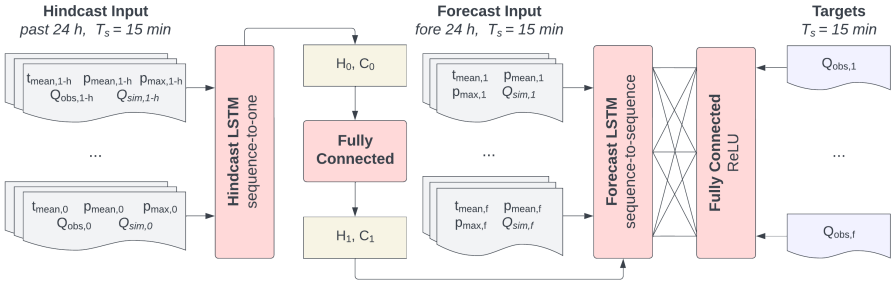

Figure 4. LSTM architecture. The optimized LSTM models incorporated the meteorological quantities p_{mean}, p_{max}, and t_{mean} in both the hindcast and forecast periods. Furthermore, the observed runoff Q_{obs} was used as a feature for the hindcast LSTM. HLSTM-PBHM exclusively incorporated results from the PBHM Q_{sim} 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.

3.2 Model performance evaluation

We utilized the five cross-folds (2013 through 2017) presented in Sect. 3.1.1 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 LSTM approach improve the overall quality of the forecasts, particularly for longer lead times? To answer this question, we first evaluated each model's (ARIMA, HLSTM-PBHM, and HLSTM) annual performance, i.e., the overall performance for each of the five previously unseen testing years. For this evaluation, we utilized three well-established metrics in hydrology, namely the NSE, the KGE, and the PBIAS. For each metric, we computed the annual average across the forecast horizon as well as the individual values corresponding to the forecast steps. Additionally, we conducted a direct comparison between both correction models, namely ARIMA and HLSTM-PBHM. This was achieved by counting the number of superior model performances for each year and lead time step. More specifically, we counted the number of times each model's predictions were closer to the observed runoff values. By normalizing the superior model performances with the number of predictions, we obtained a ratio that indicates how often each model outperformed the other. Henceforth, we will refer to this ratio as the normalized win ratio. Besides the normalized win ratio, we also investigated the overall stability of the forecasts. The forecast stability was gauged by monitoring the development of the absolute error (AE) and its variability across the entire forecast horizon. The variability was assessed by evaluating the standard deviation of the forecast errors for each forecast step. In general, a model with a high forecast stability is expected to display a relatively small standard deviation and an AE close to zero.

- How does the LSTM approach improve the forecast quality for flood events? This question was addressed by conducting a detailed investigation of each model's performance for the two largest flood events in each year. For each event, we included 12 hours before and after the maximum observed runoff into the evaluation. We then tested how well the models

https://doi.org/10.5194/egusphere-2024-1030
Preprint. Discussion started: 14 May 2024

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 the evaluation window. Additionally, we conducted a direct comparison between the ARIMA and HLSTM-PBHM correction models only considering the largest runoff values in each year. For this evaluation, we utilized the largest 5 % of the annual runoff values. This direct comparison between the correction models was done analogously to the methodology presented in the previous point.

– Does the inclusion of the PBHM's results improve the predictive skills of the LSTM? This critical question was addressed by comparing the proposed HLSTM-PBHM model to a variant that did not utilize information from the PBHM (HLSTM). More specifically, we compared differences in the annual and peak performances of both models. Additionally, we investigated the generalization capabilities of both model variants. To establish a baseline, we also evaluated the generalization capabilities of the ARIMA model. The generalization capability of each model was measured by computing the mean differences of the NSE, KGE, and PBIAS metrics obtained in the validation and testing years across all cross-folds.

4 Results

4.1 Annual model performance

4.1.1 Average performance and generalization capability

Evaluating the average annual model performances showed that all investigated model variants were able to enhance the underperforming PBHM's results. Each model's annual NSE, KGE, and PBIAS metrics, averaged over the 24-hour forecast horizon, are reported in Table 2.

The results revealed that the LSTM-based models excelled in terms of NSE and KGE. For instance, they were able to elevate the average NSE values of the PBHM from 0.19 to at least 0.87 in 2013. Even in the worst-performing year, 2017, the LSTM-based models were able to elevate the average KGE and NSE values 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 compared to the other model variants. We found that this can be attributed to the fact that our ARIMA model performed exceptionally well for forecasts that followed a clear trend or pattern. Noteworthy, 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. The significant performance gap between the PBIAS and the NSE and KGE metrics, however, suggested shortcomings of the forecasts obtained by ARIMA. The most straightforward way to identify these shortcomings was by dissecting the individual components of the KGE efficiency metric. This metric consists of three components that measure the linear correlation, the bias, and the variability between the simulated and observed 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. However, regarding the linear correlation term, we found that the LSTM forecasts significantly outperformed

https://doi.org/10.5194/egusphere-2024-1030
Preprint. Discussion started: 14 May 2024

Table 2. Average model performance comparison. Included are the annual averages of the KGE, NSE, and PBIAS metrics averaged across the entire forecast horizon for all years used for evaluation. The best values per metric and year are highlighted in **bold**.

year	PBHM			ARIMA			HLSTM-PBHM			HLSTM		
	KGE	NSE	PBIAS	KGE	NSE	PBIAS	KGE	NSE	PBIAS	KGE	NSE	PBIAS
2013	0.63	0.19	-10.68	0.84	0.69	**-0.02**	0.87	**0.88**	9.10	**0.90**	0.87	8.27
2014	0.74	0.49	-18.44	0.85	0.74	**0.03**	0.93	**0.94**	3.28	**0.95**	**0.94**	-4.54
2015	0.51	0.24	2.51	0.82	0.70	**0.07**	**0.94**	**0.91**	-1.79	0.82	0.83	-6.22
2016	0.74	0.51	-8.24	0.85	0.71	**-0.02**	0.86	0.86	5.31	**0.88**	**0.88**	-5.67
2017	0.19	-4.24	-57.66	0.56	0.02	**-0.33**	**0.87**	0.74	-8.06	**0.87**	**0.80**	2.68

those of ARIMA. According to Gupta et al. (2009), this term is influenced by the model's ability to capture the peak timing as well as the rising and falling limbs of the hydrographs.

To assess the forecast models' generalization capabilities, we computed the absolute differences (Δ) between the hydrologic metrics obtained in the validation and testing periods of each cross-fold, respectively. The resulting generalization errors are presented in Table 3.

The presented results, in general, demonstrate satisfying generalization capabilities of all model variants. Especially ARIMA, in most instances, performed exceptionally well in this regard. This was found to be particularly true for the PBIAS metric, which comes as no surprise given the exceptional performance of ARIMA for this metric. Noteworthy, also for the KGE and NSE metrics, ARIMA in most instances demonstrated superior generalization capabilities compared to the LSTM-based models. The only exception to this was found to be ARIMA's performance in 2017, which was significantly worse compared to the previous years. This in turn had a large impact on the generalization error, which resulted in 0.318 for the KGE and 0.768 for the NSE, respectively. The reason for that was the poor performance of the PBHM in 2017, upon which ARIMA heavily relied on. Interestingly, despite the PBHM's poor performance in 2017, it did not compromise the generalization capabilities of the LSTM variant that incorporated the PBHM's results. Even in this year, the HLSTM-PBHM model performed comparably well to the standalone LSTM model, HLSTM. Also for all preceding evaluation years, our results did not reveal notable differences in the generalization capabilities between the two LSTM variants.

4.1.2 Performance over lead time

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

As anticipated, both the ARIMA and LSTM models surpassed the PBHM's results across all evaluated metrics and years. ARIMA, in particular, demonstrated an exceptional performance in terms of PBIAS. Also in terms of NSE and KGE, ARIMA showed an outstanding forecast accuracy for the first forecast steps. However, this 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

https://doi.org/10.5194/egusphere-2024-1030
Preprint. Discussion started: 14 May 2024

Table 3. Generalization errors between validation and testing years for the KGE, NSE, and PBIAS metrics. Given is the absolute difference between the respective metrics. The best values per metric (i.e., the value closest to zero) and year are highlighted in **bold**.

year	ARIMA (Δ)			HLSTM-PBHM (Δ)			HLSTM (Δ)		
	KGE	NSE	PBIAS	KGE	NSE	PBIAS	KGE	NSE	PBIAS
2013	**0.024**	**0.044**	**0.096**	0.049	0.095	6.751	0.294	0.205	4.752
2014	**0.001**	0.036	**0.165**	0.004	0.028	4.333	0.015	**0.022**	4.805
2015	0.033	0.042	**0.013**	**0.029**	**0.041**	1.188	0.155	0.119	6.900
2016	**0.040**	**0.021**	**0.076**	0.109	0.072	4.346	0.086	0.045	5.044
2017	0.318	0.768	**0.367**	0.054	0.130	8.315	**0.049**	**0.089**	2.436

prediction step to 0.52 in the last. An even more significant performance decrease was observed for the NSE metric, where ARIMA achieved a value of 0.97 in the first step but merely -0.10 in the last. Compared to the forecasts obtained by ARIMA, the LSTM models generally displayed a lower accuracy in the first forecast steps. However, they were able to mostly sustain their initial accuracy across the entire forecast horizon. Even in the worst-performing year, 2017, the LSTM models were 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 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. 4.1.1.

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 not.

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 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 across the forecast horizon. The variability of the forecasts was quantified by means of the standard deviation. The results of this evaluation are presented in Fig. 6.

https://doi.org/10.5194/egusphere-2024-1030
Preprint. Discussion started: 14 May 2024

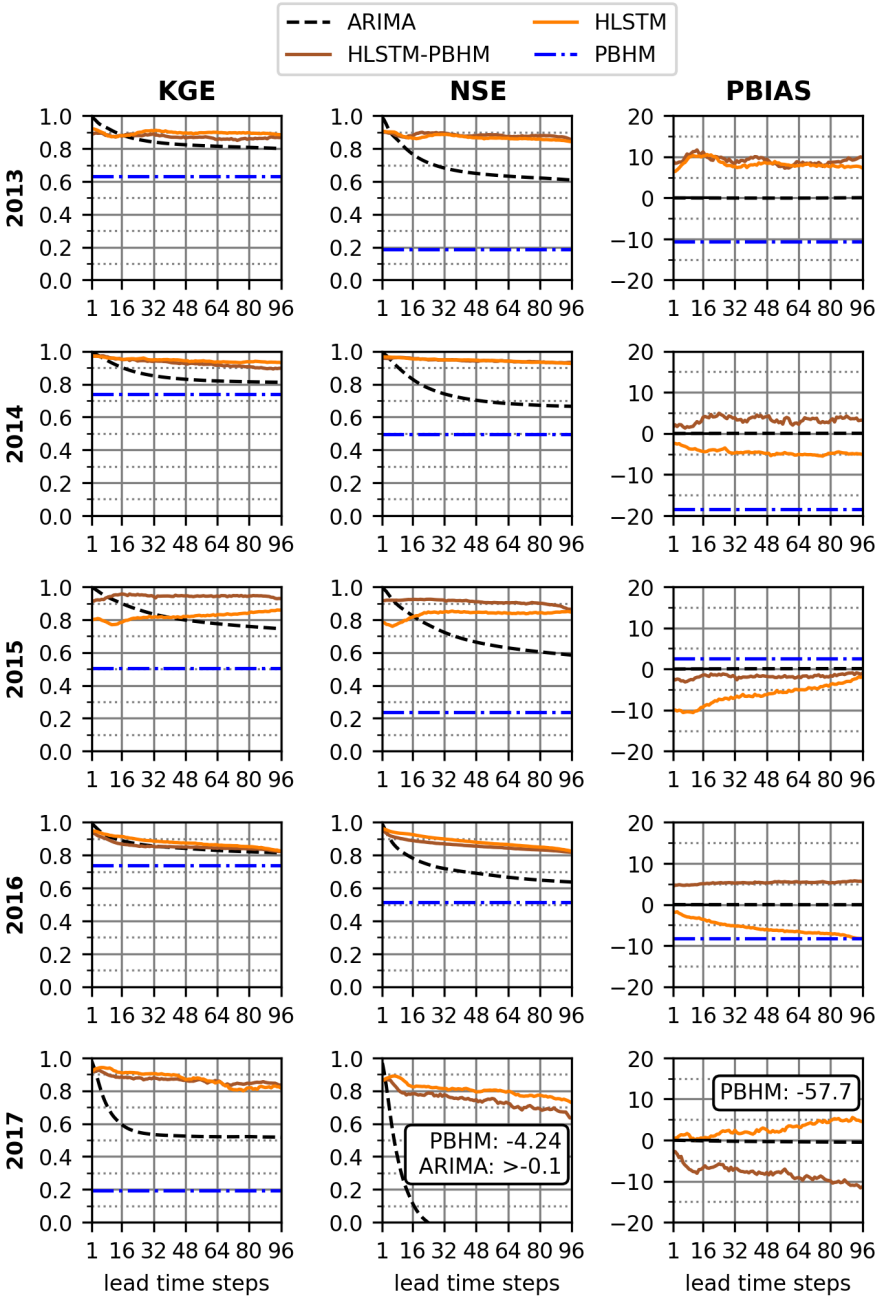

Figure 5. Development of the KGE, NSE, and PBIAS metrics over the 24-hour (96 lead time steps) forecast horizon. The evaluations include all developed model variants and all testing years.

https://doi.org/10.5194/egusphere-2024-1030
Preprint. Discussion started: 14 May 2024

When solely considering the normalized win ratio, ARIMA showed to outperform HLSTM-PBHM more often than not. Especially, in the first forecast step 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. 4.1.1). 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 times. This indicates that ARIMA in certain instances produced considerably worse forecasts compared to the LSTM. In this regard, HLSTM-PBHM surpassed ARIMA on average after four forecast steps (1.00 hours).

4.2 Performance for elevated river runoff

4.2.1 Peak timing and magnitude

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 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 4.

Upon initial inspection, the presented results highlight the deficiencies of the PBHM in capturing both the peak magnitude and its timing. Especially, the substantial timing errors suggest shortcomings of the model in adequately depicting the characteristics of the hydrographs. In terms of magnitude error, the PBHM exhibited a median magnitude error of -44.7 %, 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 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 PBHM, its magnitude errors were still large.

In terms of timing errors, 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 timing errors in the forecasts. More specifically, both LSTM variants achieved a median absolute timing error of two time steps, equivalent to 30 minutes.

https://doi.org/10.5194/egusphere-2024-1030
Preprint. Discussion started: 14 May 2024

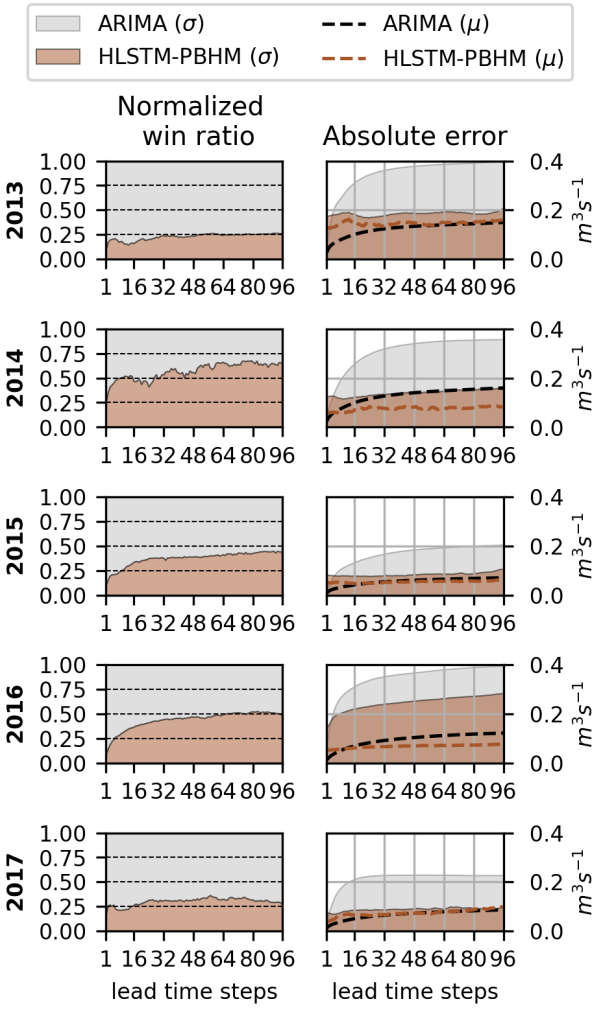

Figure 6. ARIMA and HLSTM-PBHM forecast comparison for all forecasts per year for the 24-hour forecast horizon (96 time steps). (Left) Normalized win ratio (i.e, ratio of superior model forecasts). (Right) Propagation of the absolute error. Shown are the mean μ and the standard deviation σ of the absolute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

4.2.2 ARIMA and HLSTM-PBHM comparison for elevated river runoff

Similar to the results presented in Sect. 4.1.3, Fig. 7 shows the normalized win ratio but only evaluated for the largest 5 % of the annual runoff. Furthermore, the propagation of the absolute error and its variability are shown.

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 of the mean absolute error and its standard deviation. Except for 2015, the LSTM's absolute error for longer lead times was

https://doi.org/10.5194/egusphere-2024-1030
Preprint. Discussion started: 14 May 2024

Table 4. Comparison of the median peak magnitude e_{peak} (in percent) and timing errors 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		HLSTM-PBHM		HLSTM	
		(m^3s^{-1})	e_{peak} (%)	Δt	e_{peak} (%)	Δt	e_{peak} (%)	Δt	e_{peak} (%)	Δt
2013	1st	15.00	-90.3	31	-75.3	47	-44.7	29	**-16.2**	**0**
	2nd	10.02	+13.3	20	**-3.5**	19	-49.8	8	+17.4	**3**
2014	1st	7.27	+3.5	18	**+2.1**	18	-33.5	**-1**	-46.1	7
	2nd	6.23	+23.3	16	+21.3	16	**-20.0**	**-2**	-37.6	2
2015	1st	5.85	-62.5	36	-46.9	36	-22.5	**0**	**-15.9**	-2
	2nd	3.33	**+4.7**	49	+6.2	48	-9.8	**2**	-18.9	2
2016	1st	17.94	-73.6	26	-47.4	47	-23.8	2	**+23.4**	**1**
	2nd	9.99	-39.4	-96	-33.9	-89	-67.5	4	**-25.4**	**2**
2017	1st	9.21	-49.9	25	**-43.1**	37	-56.4	63	-62.0	**3**
	2nd	7.37	-63.1	28	-44.7	29	**-16.2**	**0**	-47.6	57

almost half of ARIMA's errors. A similar trend was observed for the standard deviations. Noteworthy, also for elevated runoff, ARIMA's forecast accuracy 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.

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, we developed an LSTM model (HLSTM-PBHM) that was largely inspired by the hindcast-forecast architecture presented by Nevo et al. (2022). This specific architecture was selected 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 forecasts and analyses, potentially enabling the model to learn from their differences. Furthermore, with the here presented cross-validation strategy, we established a framework for a seamless continuous improvement of the model as new observational data become available. To showcase the proposed model's effectiveness, we rigorously compared its forecasting capabilities to those of a more 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 for longer lead times and flood event runoff, both being recognized weaknesses of ARIMA models.

When comparing the forecasts obtained by both correction models, ARIMA and HLSTM-PBHM, we observed that both had their advantages and disadvantages. ARIMA generally showed a very high accuracy in the first forecast steps. However,

https://doi.org/10.5194/egusphere-2024-1030
Preprint. Discussion started: 14 May 2024

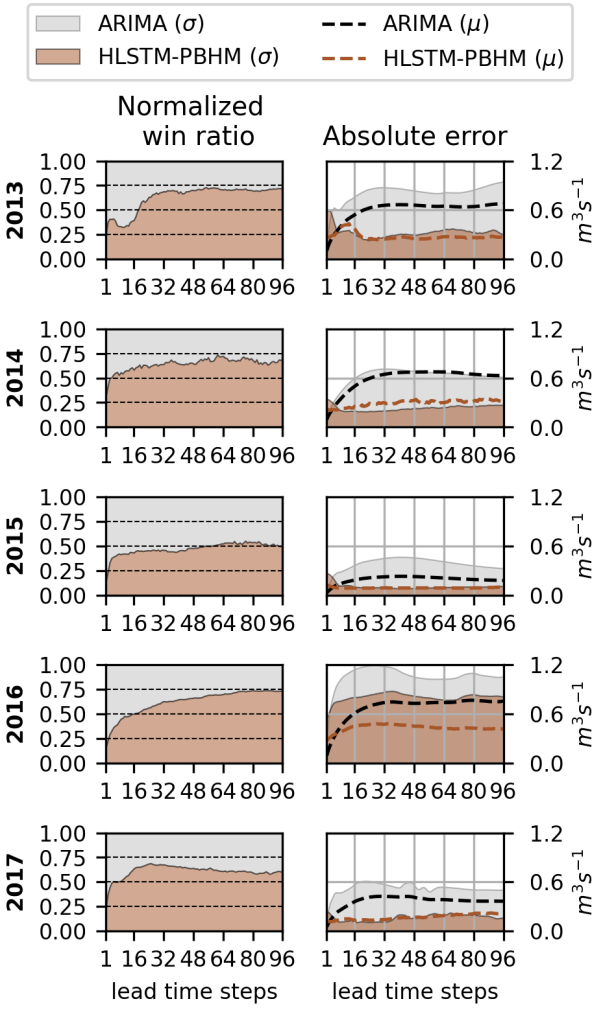

Figure 7. ARIMA and HLSTM-PBHM forecast comparison for the largest 5 % of the annual runoff values for the 24-hour forecast horizon (96 time steps). (Left) Normalized win ratio (i.e, ratio of superior model forecasts). (Right) Propagation of the absolute error. Shown are the mean μ and the standard deviation σ of the absolute errors.

this initial accuracy 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 exhibited a larger error in the first steps but it was able to mostly sustain its initial accuracy across the 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. Interestingly, ARIMA

https://doi.org/10.5194/egusphere-2024-1030
Preprint. Discussion started: 14 May 2024

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 most often during base flow conditions.

When considering the forecast skills at flood event runoff, the LSTM outperformed ARIMA. This was indicated by both the KGE's Gupta et al. (2009) correlation term and also the obtained errors at selected flood events. Although both the LSTM and ARIMA models were not 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. 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 by Liu et al. (2015). Overall we found that our LSTM model outperformed ARIMA in all aspects we consider relevant for operational flood forecasting, i.e., a more accurate representation of flood event runoff and more reliable forecasts for longer lead times.

Reflecting on the remarkable capabilities of LSTM models in predicting river runoff (e.g., Kratzert et al., 2019) inevitably prompts the question 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'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 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.

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., McMahon and Peel, 2019). 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., Natel de Moura et al., 2022). (III) Limited availability of system states: The information of system states in the catchment is limited when employing solely data-driven models. However, many operational forecasting systems rely on information of the system states, e.g., 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 more complex forecasting chains.

https://doi.org/10.5194/egusphere-2024-1030
Preprint. Discussion started: 14 May 2024

Although the here presented LSTM models already achieved a comparably high forecast accuracy, there a potential exists for future enhancements. Firstly, the pre-processing phase can be intensified. Specifically, a more careful approach to feature engineering could increase the model's quality by providing more relevant and informative features included in training. Secondly, 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.

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's results (HLSTM). The main findings of this study can be summarized as follows:

- Both correction methods (ARIMA and HLSTM-PBHM) were able to elevate the forecast accuracy of the exiting PBHM.
- ARIMA achieved a particularly high forecast accuracy in the first forecast steps. However, this initial accuracy showed to decline quickly with increasing lead time. Contrary to that, the LSTM models exhibited a larger error in the first steps, but they were able to mostly sustain their initial accuracy across the forecast horizon.
- Both, ARIMA and HLSTM-PBHM, displayed shortcomings in accurately predicting the magnitude of the largest flood events. However, in contrast to ARIMA, HLSTM-PBHM was able to accurately predict the timing of the maximum peak runoff.
- We did not find strong evidence of whether the inclusion of the PBHM's results benefited the accuracy of the LSTM. Both models, the one that utilized the PBHM's results and the one that did not, yielded comparably accurate forecasts.

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 PBHM is underperforming.

Appendix A: Evaluation metrics

A1 Nash-Sutcliffe efficiency (NSE)

The Nash-Sutcliffe 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:

https://doi.org/10.5194/egusphere-2024-1030
Preprint. Discussion started: 14 May 2024

$$NSE = 1 - \frac{\sum_{t=1}^{N}(Q_{Obs,t} - Q_{Sim,t})^2}{\sum_{t=1}^{N}(Q_{Obs,t} - \overline{Q}_{Obs})^2} \tag{A1}$$

where $Q_{Obs,t}$ 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.

A2 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:

$$KGE = 1 - \sqrt{(r-1)^2 + (\beta - 1) + (\gamma - 1)^2} \tag{A2}$$

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 $\mu_{Sim,t}/\mu_{Obs,t}$ and γ is the variability term, which is computed from the standard deviations and the mean values as $\frac{\sigma_{Sim,t}/\mu_{Sim,t}}{\sigma_{Obs,t}/\mu_{Obs,t}}$. The KGE is bound between 1 and $-\infty$, with 1 indicating perfect model predictions.

A3 Percent 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:

$$PBIAS = \frac{\sum_{t=1}^{N}(Q_{Obs,t} - Q_{Sim,t})}{\sum_{t=1}^{N} Q_{Obs,t}} \cdot 100 \tag{A3}$$

where $Q_{Obs,t}$ and $Q_{Sim,t}$ 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 B: Loss function

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:

$$f_{obj} = 2 - KGE - NSE \tag{B1}$$

https://doi.org/10.5194/egusphere-2024-1030
Preprint. Discussion started: 14 May 2024

Table C1. Hyperparameter search space and final parameter set for the LSTM models

Hyperparameter	Search space		HLSTM-PBHM	HLSTM
	Min.	Max.		
N. of LSTM units	24	96	96	46
Batch size*			4000	4000
Initial learning rate	1e-3	1e-2	0.0010	0.00223
Retrain epochs*			5	5
Dropout rate	0.01	0.5	0.370	0.418

*is kept constant to speed up model training

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 a perfect fit by the model. The resulting loss function can be written as:

$$Loss = 2 - normKGE - normNSE \tag{B2}$$

Appendix C: Auxiliary information for LSTM hyperparameter tuning

Table C1 shows the LSTM hyperparameters subjected to optimization, their search space as well as their final values for both model variants.

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: The runoff was provided by the 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) Meteorological data: The meteorological 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**.

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

https://doi.org/10.5194/egusphere-2024-1030
Preprint. Discussion started: 14 May 2024

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

Acknowledgements. We express our gratitude to the 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.

https://doi.org/10.5194/egusphere-2024-1030
Preprint. Discussion started: 14 May 2024

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.

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.

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.

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.

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 Real-time Flood Forecast Correction: 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.

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.

https://doi.org/10.5194/egusphere-2024-1030
Preprint. Discussion started: 14 May 2024

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., 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, 2019.

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.

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.

McMahon, T. A. and Peel, M. C.: Uncertainty in stage–discharge rating curves: application to Australian Hydrologic Reference Stations data, Hydrological sciences journal, 64, 255–275, 2019.

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.

Natel de Moura, C., Seibert, J., and Detzel, D. H. M.: Evaluating the long short-term memory (LSTM) network for discharge prediction under changing climate conditions, Hydrology Research, 53, 657–667, 2022.

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.

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

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.