Redevelopment of LobithNN

On working methods, model choices and model quality

Redevelopment of LobithNN

On working methods, model choices and model quality

Author(s)

Jurian Beunk

Jing Deng

Contents

1 Introduction

In RWsOS-Rivieren a data-driven approach is used for discharge forecasting for the first two days at Lobith. The first model developed for this task was a multi-linear regression model, driven by upstream water level and precipitation measurements. This model was named LobithW and operates on a daily timescale.

In 2023 Deltares developed and implemented a new model, which aimed to model short-term discharge at an hourly timescale and with improved quality over LobithW (Appendix A). This model was named LobithNN and was based on a deep-learning approach using Long Short-Term Memory networks (LSTMs) (Hochreiter & Schmidhuber, 1997).

After operationalizing LobithNN, three issues were identified:

1. Model performance decreased in operation. This issue was later attributed to the fact that the operational data (Pegelonline) was different from the training data (BfG).
2. Large errors were observed in the first forecast timestep.
3. The model failed regularly due to missing data.

As a result, Rijkswaterstaat has asked to address these issues and improve the model. The objective is therefore to develop a model that:

- has skill over LobithW in the operational setting.
- initializes well at the first timestep.
- is less sensitive to missing data.

The memo is structured as follows: Section 2 describes the data, testcases, model architectures, model performance evaluation metrics, and the eXplainable Artificial Intelligence (XAI) technique used in this project. Section 3 presents the best model setup obtained through experimentation with testcases, discusses the quality and operational reliability of this model, summarizes the results for different testcases, and provides example results of the XAI technique in use. Section 4 outlines the limitations of this project and proposes future works. The memo concluded with a summary in Section 5.

2 Methods

2.1 Data

In this project, we design a data-driven model to forecast hourly discharge at Lobith with lead times up to 48 hours ahead. We experiment with upstream water levels, historical discharge at Lobith, and precipitations as inputs to the model.

2.1.1 Discharge at Lobith

The hourly discharge observations at Lobith were obtained from the Rijkswaterstaat Waterinfo website[1]. The raw data presented two issues. Firstly, before approximately 1997, the hourly data were generated by interpolating daily measurements, which is not good data to train an hourly model. Therefore, we decided to use data after 1997. Secondly, there are anomalous peaks and valleys in the original raw data. We developed a method based on the function find_peaks from scipy.signal python library to detect and filter out these anomalies (see Figure 2-1 as an example). The pre-processed data resulted in better model performance compared to the raw data; thus, the pre-processed data was used to train the model.

Although it is beyond the scope of this project, we recommend investigating the causes behind the anomalous peaks and valleys in the measurements.

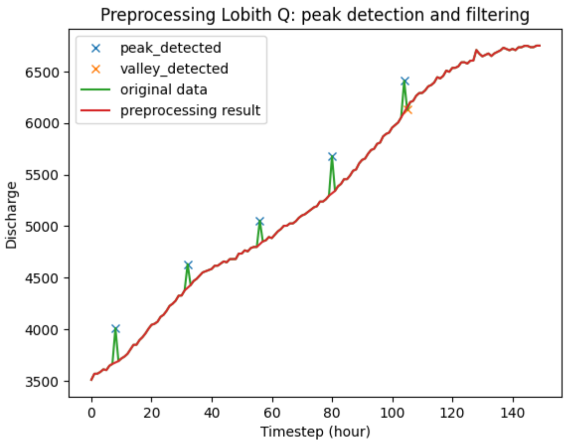


Figure 2-1 Preprocessing of discharge observations at Lobith: peak and valley detection and filtering.

2.1.2 Upstream water levels

Two data sources are available for the hourly water levels from upstream stations: BfG and Matroos. BfG data is post-processed from measurements and covers a longer period, from 1980 to 2021. However, it contains missing values before the 1990s. Matroos data is used in RWsOS operational systems but has a much shorter period of available data, from September 2009 until now, but has quite some missing values.

We initially trained the model using Matroos data, but the limited training samples led to worse model performance when compared to using BfG data in training. Therefore, we decided to use BfG data for model training.

[1] https://waterinfo.rws.nl/#/nav/index

Since the operational model will use Matroos data as input, it was necessary to evaluate the model trained on BfG data using Matroos data. The results confirmed that it is feasible to train the model on BfG data and use Matroos data for inference.

2.1.3 Precipitation

Precipitation was taken from the ERA5 dataset and averaged across potentially relevant sub-basins of the Rhine. The sub-basins included: Lower Rhine, Middle Rhine, Sies, Lippe, Ruhr, Erft and Wupper.

2.2 Testcases

Based on the identified issues, various testcases were developed to improve the model. The original testcases, as outlined in the memo “*Plan van aanpak voortzetting Lobith-NN”* (Appendix A), included:

- Testcase 1: less upstream stations.
- Testcase 2: training on time-differenced discharge.
- Testcase 3: different temporal resolution (1-hourly or 3-hourly).
- Testcase 4: adding precipitation as input.
- Testcase 5: trying a CNN model structure.

However, during the experimentation process, we further defined our scope. As a result, we conducted testcase 1, 2, 4, and an additional testcase – testing different look back windows.

For each individual test case, we conducted many experiments. Experiments were designed by following an iterative approach, evaluating many combinations of variables (such as input stations). In addition, SHAP analysis was used during model development to assist in finding the optimal settings for experiments (more about SHAP in Section 2.5).

For testcase 1, based on the reliability of BfG data from upstream stations, we selected eight upstream stations: Rees, Wesel, Düsseldorf, Köln, Andernach, Koblenz, Kaub, and Maxau. We further refined the selection based on system understanding and feature importance analysis using SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017). For example, Rees and Wesel are close to each other and have highly correlated water levels, and Maxau's water level did not contribute significantly to the results. Ultimately, we limited the stations to six: Wesel, Düsseldorf, Köln, Andernach, Koblenz, and Kaub.

For testcase 2, we experimented with using time-differenced discharge at Lobith (*dQ*, i.e., the discharge difference between two consecutive hours) as the target instead of discharge (*Q*). Predicting value changes is a common approach in machine learning. This test case aimed to determine if predicting *dQ* would improve performance.

For testcase 4, we experimented with adding observed and/or forecasted precipitation from the nearest catchments upstream of Lobith. The assumption was that local precipitation might contribute to short-term high discharge, thus improving model performance.

Our model requires inputs from the past x days, known as the look-back window. For the additional testcase - different look back window, we experimented with look-back windows of 2, 3, 4, 5, 6, 8, 10, 15, and 20 days with the reduced number of stations.

2.3 Model architecture

We designed a single layer LSTM model architecture for this project (Figure 2-2). This architecture uses a sequence of LSTM units to process historical observations of water levels at upstream stations and discharges at Lobith. The output from the LSTM layer is then fed into a Dense layer, which generates the final output of 48 predictions simultaneously. This model architecture is used for all testcases except for the one that incorporates forecasted precipitation in testcase 4.

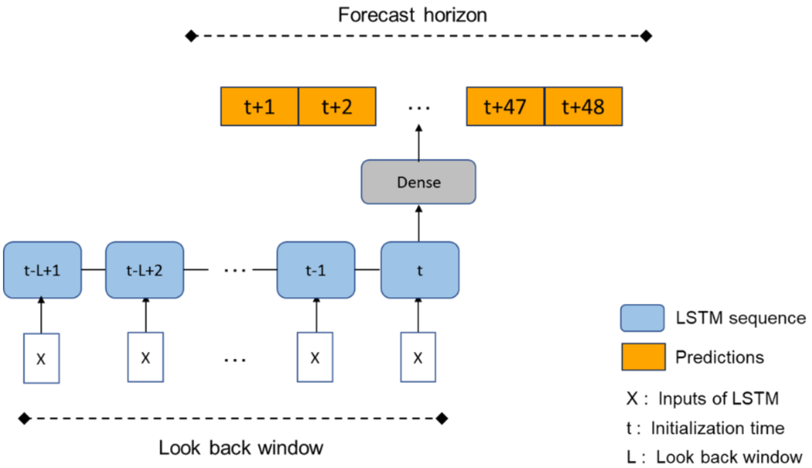


Figure 2-2 Schematic illustration of the basic model architecture used in this project.

In testcase 4, the basic model architecture is modified to be able to incorporate forecasted precipitation. The modified architecture is shown in Figure 2-3. The modified model architecture utilizes two LSTM models, with one LSTM (LSTM-1) processing historical observation data (i.e., historical observation of water levels at upstream stations, discharge at Lobith, and/or catchment-averaged precipitation) and another LSTM (LSTM-2) processing forecast data (i.e., forecasted catchment-averaged precipitation). The outputs of each step in LSTM-2 are then fed into the Dense layer, which produces the final output of 48 predictions simultaneously.

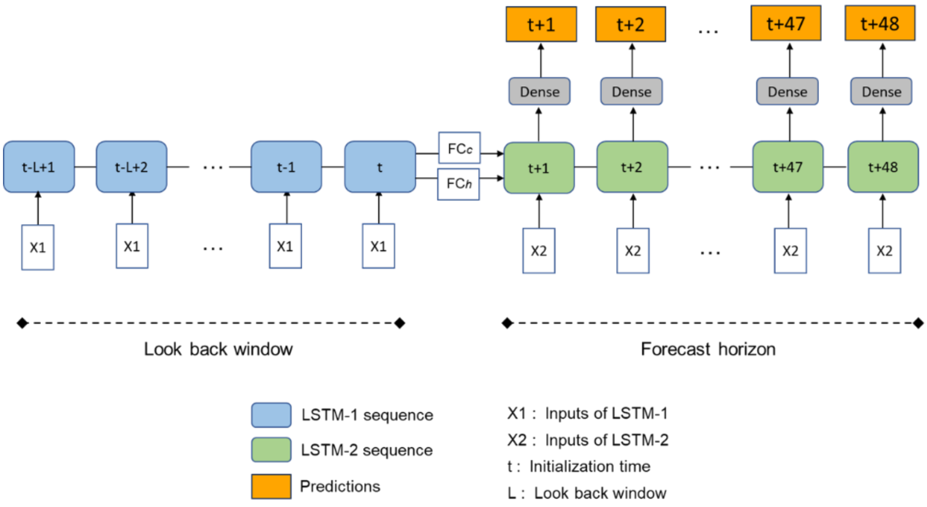


Figure 2-3 Schematic illustration of the modified model architecture to incorporate forecasted precipitation in testcase 4.

2.4 Evaluation metrics

We use mean absolute error (MAE), mean absolute percentage error (MAPE), and bias as metrics to evaluate the model performance across different lead times. MAE, in the same unit as the target (m^3/s), offers a straightforward interpretation and is commonly used in deterministic forecast. The discharge at Lobith involves large discharge ranges across different seasons. MAPE, expressed in unit 100%, normalizes errors by expressing them as a percentage of the actual discharge, which facilitates a more equitable comparison of forecasting performance across seasons. Bias can give us an idea if the model is structurally underestimating or overestimating discharges.
Furthermore, we evaluate the new LobithNN model performance in hindcast settings compared to benchmark model LobithW and the previous version of LobithNN using skill score based on MAE and MAPE (see Equation (1) and (2)). The skill scores range from negative infinity to 1. If the skill score is above zero, it indicates that the new LobithNN model has more skill compared to the benchmark. If the skill score is below zero, it indicates that the benchmark has more skill than the new LobithNN.

$$Skillscore\ (mae) = 1 - \frac{mae}{mae_{reference}} \qquad (1)$$

$$Skillscore\ (mape) = 1 - \frac{mape}{mape_{reference}} \qquad (2)$$

2.5 eXplainable Artificial Intelligence (XAI)

The success of machine learning (ML) models, especially deep learning (DL) models, is due to their advanced algorithms and large parameter spaces. However, their complexity can make them appear as "black box", obscuring their decision-making processes. Understanding the reasons behind model predictions is crucial for ensuring fairness, safety, and trust, especially in high-risk situations or when addressing biases.

In this project, we employed SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017), an eXplainable Artificial Intelligence (XAI) technique, to address the “black box” issue. We used SHAP analysis to study feature importance, which helped us identify the most critical upstream stations and therefore reduce the redundancy in model inputs. Additionally, SHAP analysis allowed us to verify whether the model’s behavior aligns with our understanding of the physical system.

3 Results

This chapter summarizes the performance of the newly developed model, i.e., the best model setup obtained through experimentation with testcases, while comparing against the currently still operational LobithW model and the previous version of LobithNN model, still operational until release RWsOS 2024.05 (thereafter referred to as Old LobithNN). The performance is quantified using evaluation metrics described in Section 2.4. Furthermore, this chapter also summarizes the results for different testcases, and provides example results of the XAI technique in use.

Additional background figures are provided in Appendix C.

3.1 The best model

The best model found in the research study was a relatively simple one. Its inputs are purely historical and a combination of discharge at Lobith and upstream measurements of water level (Figure 3-1, Figure 3-2). The upstream stations are Wesel, Dusseldorf, Koln, Andernach, Koblenz and Kaub. The history that are provided is ten days.

In this section, a detailed overview is provided of the quality of the model compared to LobithW and the previous implementation of LobithNN.

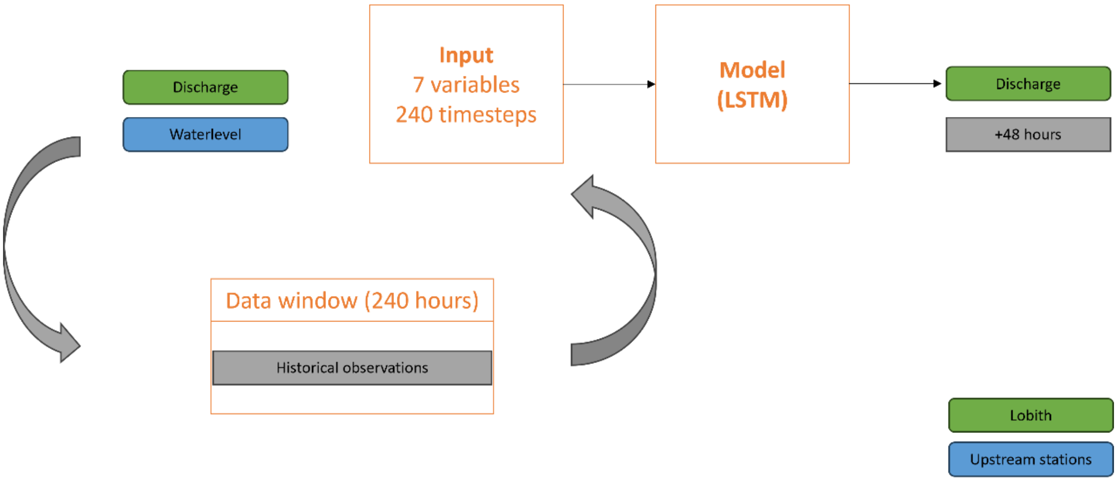


Figure 3-1 Model flow of the best model.

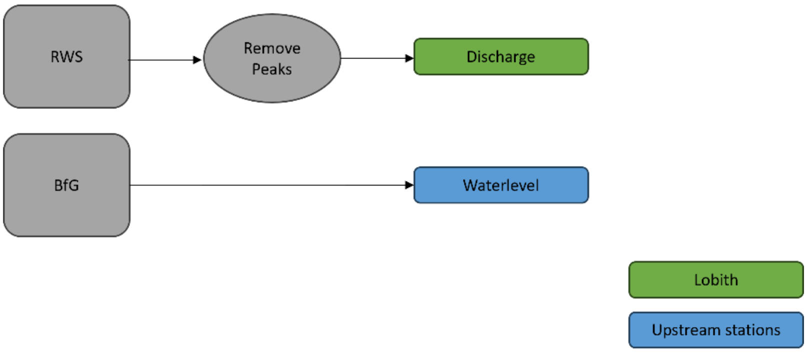


Figure 3-2 Data flow of the best model.

3.1.1 What is the quality of the best model compared to LobithW?

Forecast results of the new LobithNN were compared against forecasts results of LobithW over the period of 2019-2021.

Figure 3-3 shows that the new model outperforms LobithW across all lead times and in terms of the MAE and the MAPE.

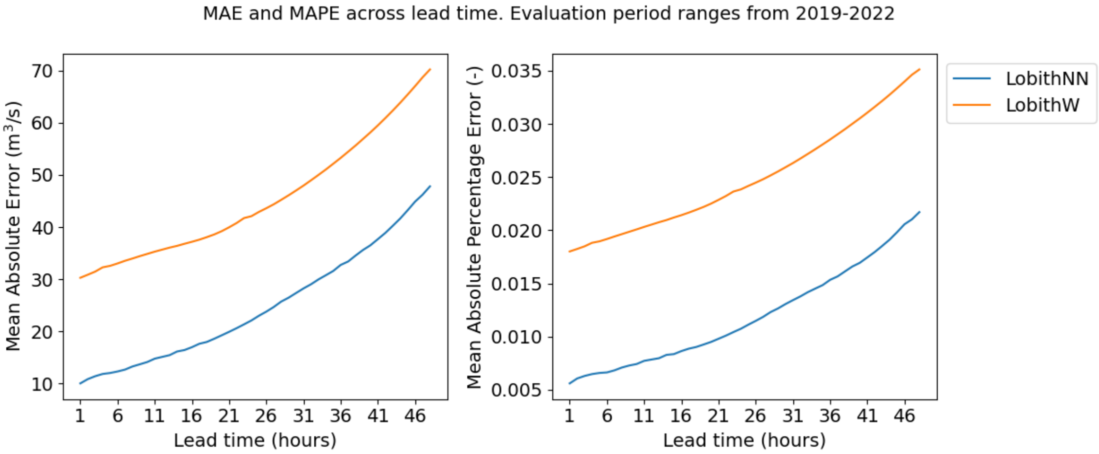


Figure 3-3 The mean absolute error (MAE, left) and the mean absolute percentage error (MAPE, right) for the best model (LobithNN) and LobithW, between 2019 and 2021.

The performance was also assessed for different ranges of discharge. The MAPE was computed for each model and are presented in the Appendix C.

Based on these scores, a skill score was computed, effectively comparing the models. Figure 3-4 shows the skill-score based on the MAE of the new model compared to LobithW across different discharge domains. Positive values (blues) indicate skill of LobithNN over LobithW. Negative values indicate skill of LobithW over LobithNN.

For discharges below 4000 m^3/s, LobithNN always outperforms LobithW. For discharges between 4000-6000 m^3/s, neither model is clearly outperforming the other and the skill scores are within the range (-0.4, 04), indicating that the MAPE is never more than 40% better or worse. For discharges above 6000 m^3/s, LobithW has skill over LobithNN. It should be noted that the sample size in this upper discharge bin was limited, only covering three high-flow events.

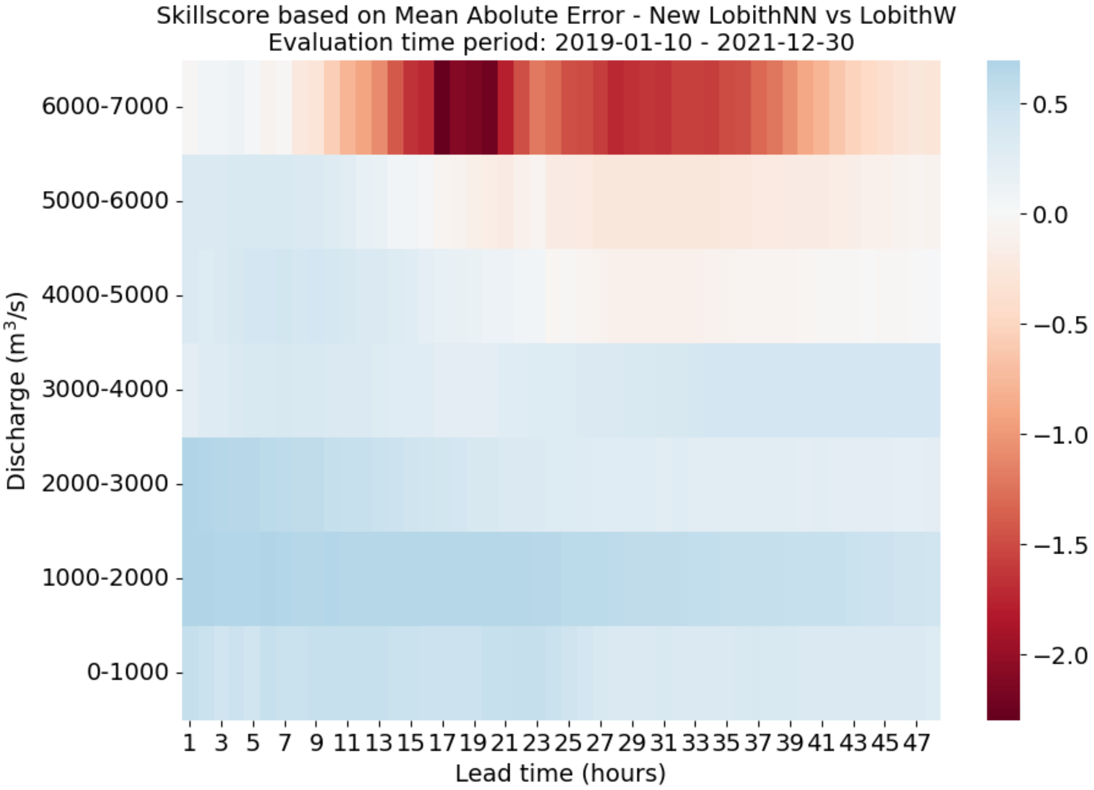


Figure 3-4 Heatmap showing the skill scores based on the mean absolute error for LobithNN and LobithW. Positive values (blues) indicate skill of LobithNN over LobithW. Negative values (reds) indicate skill of LobithW over LobithNN.

3.1.2 Quality of the model compared to the Old LobithNN

A similar comparison was made with the Old LobithNN. Results are presented in an analogous manner. Because the Old LobithNN was trained on the full training period, model results could not be inferred from that same period. To be able to make a comparison, we downloaded more recent data (outside of the training period) from the Matroos database, and predictions were made with both the old and the new version of LobithNN. The period ranged from January 2023 until May 2024.

Figure 3-5 shows that the new model outperforms the Old LobithNN across all lead times and in terms of the MAE and the MAPE. Especially at the first forecast timestep, the new model has a lower error (MAE of +/- 11 m^3/s, compared to 22 m^3/s).

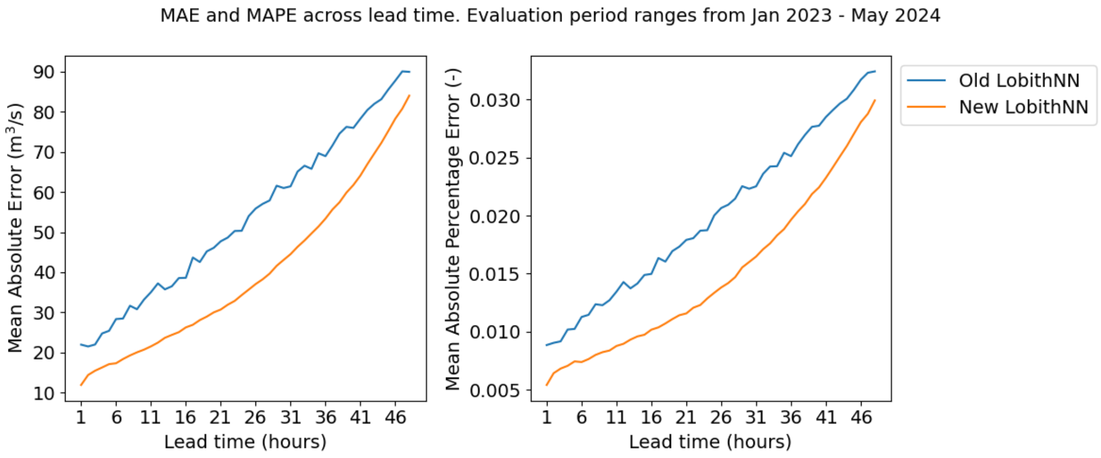


Figure 3-5 The mean absolute error (MAE, left) and the mean absolute percentage error (MAPE, right) for the best model (New LobithNN) and the Old LobithNN, between 2023 and May 2024.

Figure 3-6 provides an overview of the skill scores across lead time and across different ranges of discharge values. For nearly all combinations of lead time and discharge range, the new version of LobithNN outperforms the Old LobithNN. Again, the difference at the first forecast timestep is recognized. At the final few timesteps, and in low and high discharge ranges, the Old LobithNN sometimes outperformed the new version.

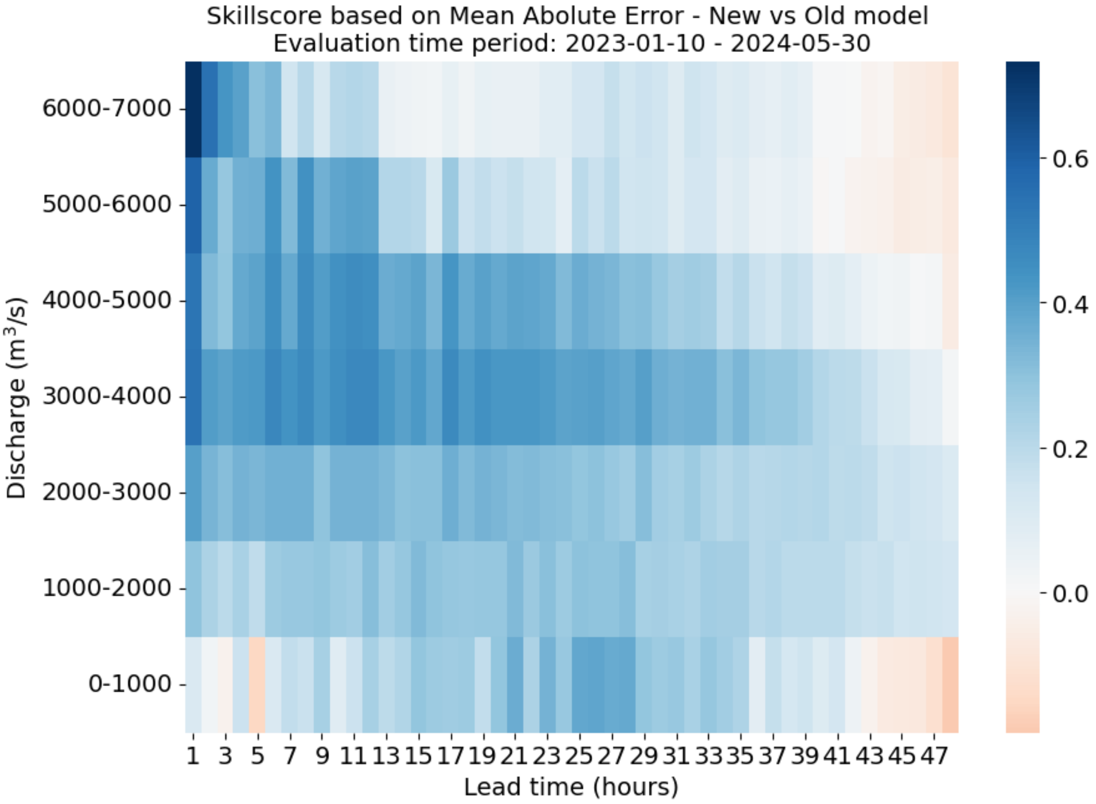


Figure 3-6 Heatmap showing the skill scores based on the mean absolute error for the new LobithNN and the Old LobithNN. Positive values (blues) indicate skill of the new version over the Old LobithNN, while negative values (reds) indicate skill of the Old LobithNN over the new LobithNN.

3.1.3 Operational reliability

LSTM-based deep-learning models do not allow for missing data. This is also the case for Multi-linear regression (MLR) model like LobithW. However, to maximize operational reliability, station reliability was considered as a constraint in the input feature selection. In additional the number of input stations was minimized. After investigating the operational archive Matroos, it was found that stations on the tributaries were less reliable than stations on the main river channel and it was chosen to focus on using these stations.

In addition to minimizing the number of input stations and avoid using the less reliable stations, simple interpolation and extrapolation of input data is applied in the operational setting. Currently input data will be interpolated for a maximum of 25 hours and extrapolated for a maximum of 6 hours.

During the project, it was noted that the currently operational LobithW has more extensive settings for both inter- and extrapolation. This may cause users to think LobithW is a more reliable model than LobithNN. It is advisable to investigate trade-off between model performance and at a later stage.

3.2 Testcase results

In this section, we summarize the results from the testcases. The quality of models was assessed using the metrics presented in the Methods (Section 2). A more extensive description of the testcases is provided in Appendix D.

In Testcase 1, eight upstream stations were initially chosen based on reliability. Removing Rees and Maxau, improved model performance due to their redundancy and minimal contribution. This resulted in the use of six stations: Wesel, Düsseldorf, Köln, Andernach, Koblenz, and Kaub. Additionally, experimenting with various look-back windows (LBWs) showed that a 10-day LBW offered slightly better performance than a 4-day LBW.

In Testcase 2, training on time-differenced discharge did not improve performance, indicating the model's struggle to learn input-output relationships.

In Testcase 4, adding precipitation data did not enhance the model, probably because this information is already reflected in the upstream water level measurements as a result of rainfall-runoff process, and adding precipitation would instead introduce confusion and redundancy to the model.

3.3 SHAP analysis

Feature importance analysis

Based on the reliability of BfG data from upstream stations, we selected eight upstream stations. To further refine the input features, we conducted SHAP analysis to study feature importance. The SHAP results are presented in Figure 3-7. We can interpret the SHAP values as levels of contribution. From the figure, it is evident that the historical discharge at Lobith (lobith_q) contributes the most for all lead times compared to other features. Among the upstream stations, the water level from Maxau station (maxau_h) contributes the least across all lead times. Therefore, it is expected that removing the Maxau station from the input will not affect the model performance.

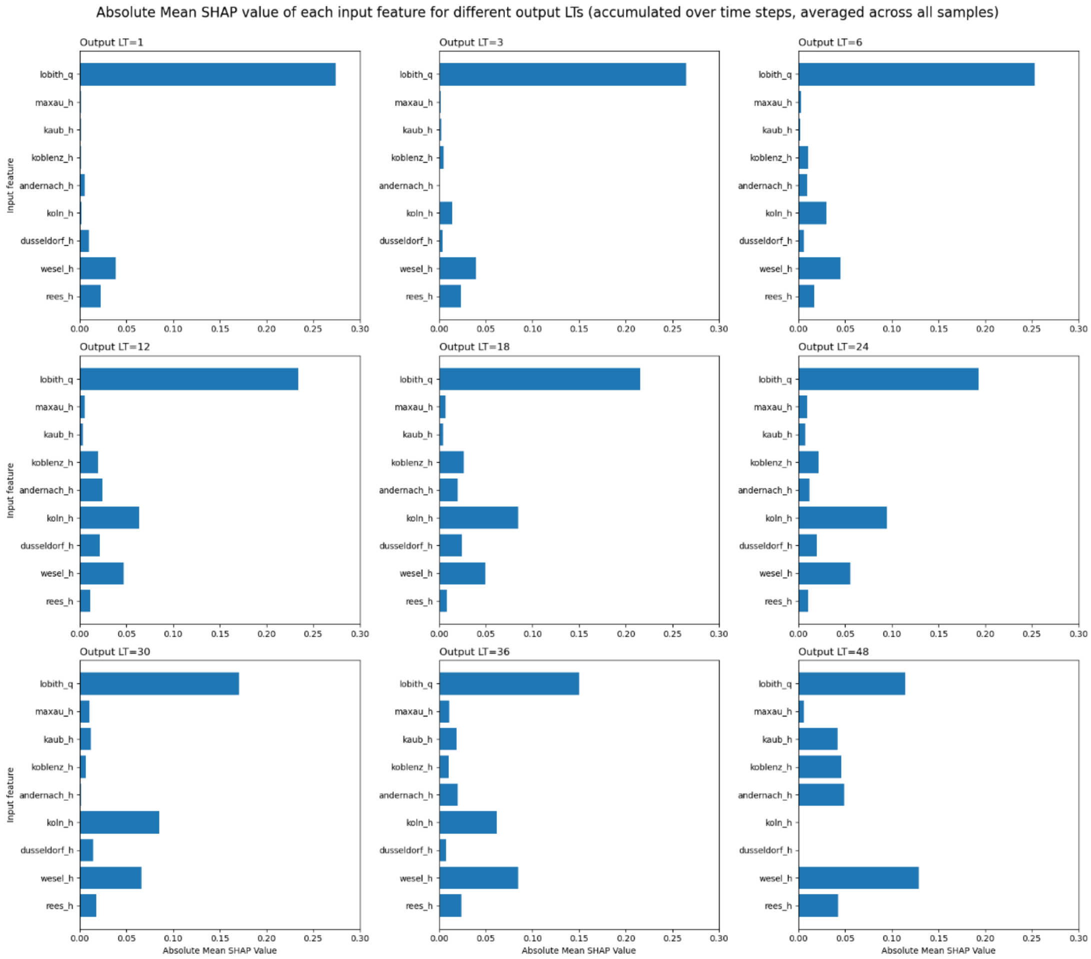


Figure 3-7 SHAP analysis for the experiments that have eight upstream stations and historical discharge at Lobith as inputs.

Analyze model behavior
We conducted SHAP analysis for our final model to check whether the model behavior aligns with our understanding of the physical system.

From the results (Figure 3-8), we observe that the contributions from upstream stations change as lead time increases. The water level from Wesel starts to contribute at an earlier lead time (starting from LT=1) because Wesel is the closest station to Lobith. Therefore, its contribution to the discharge at Lobith is the earliest. After LT=5, we observe increasing contributions from Koln. Koln and Dusseldorf are the second closest stations to Lobith. For longer lead times (after LT=30), we see increased contributions from more upstream stations (Andernach, Koblenz, and Kaub). The contributions from upstream stations change along with lead time, and the order of change aligns with the flow travel time from upstream stations to Lobith.

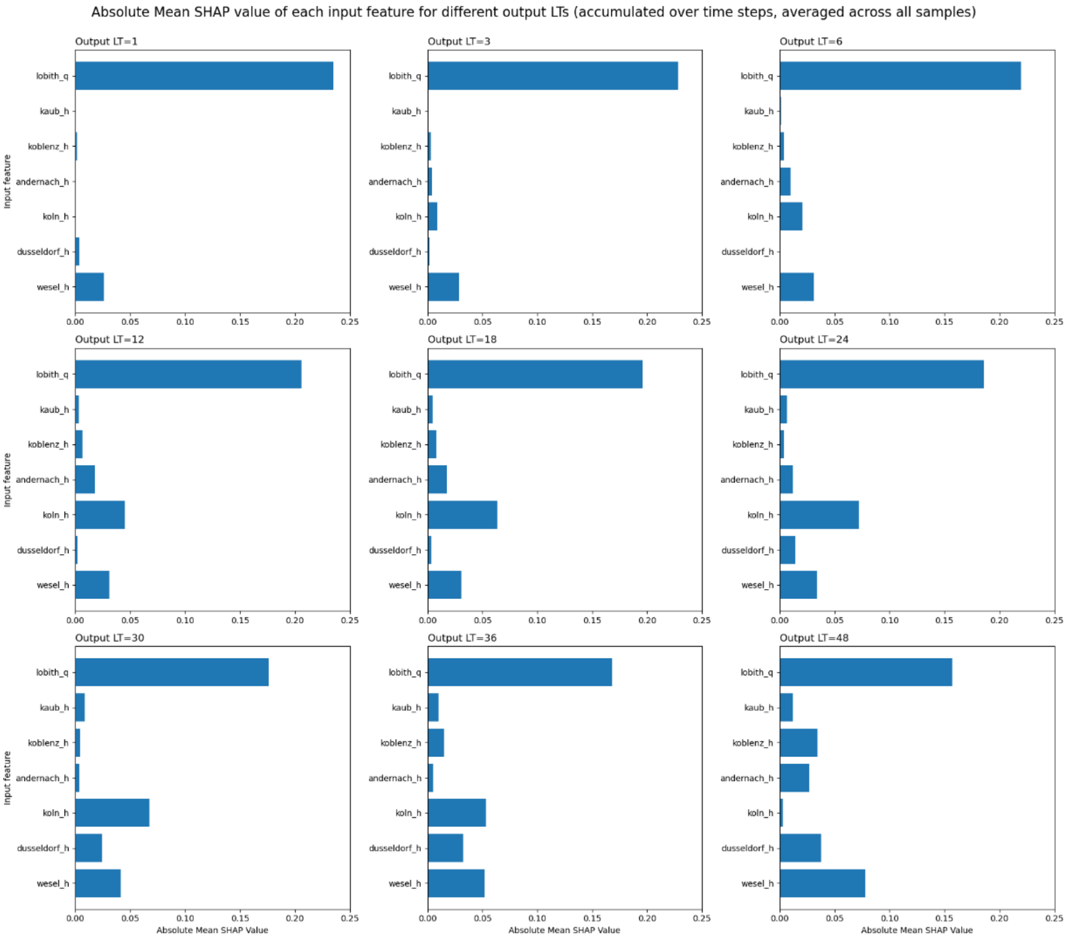


Figure 3-8 SHAP analysis for the final model (six upstream stations and historical discharge at Lobith as inputs).

3.4 Summary

In this project, a deep-learning model was developed to forecast discharge at Lobith 48 hours into the future. The model has 7 inputs (historical measurements of discharge at Lobith and water level at Wesel, Dusseldorf, Koln, Andernach, Koblenz and Kaub. The model has an hourly timestep.

- **Model performance**: the model outperforms LobithW across all lead times (MAE of +/- 48 m^3/s compared to +/- 70 m^3/s at +48 hours between 2019-2022).
- **Model performance:** between 2019 and 2022, for discharges below 4000 m^3/s, the model always outperforms LobithW. For discharges between 4000-6000 m^3/s, LobithNN performs generally better in the first day and LobithW performs better in the second day (MAE-Skill score +/- -0.2).
- **Model performance**: the model initializes well at the first timestep, resolving the issue with the previous implementation of LobithNN. The error at the first timestep is +/-12m^3/s compared to 22m^3/s with the previous version of LobithNN.
- **Operational reliability**: the model is dependent on seven input stations, which is less than the previous version of LobithNN (14 input stations) or LobithW (21 input stations). In addition, the new model only relies on stations in the main channel of the Rhine, which are more reliable. As a result of these two factors, the model is more reliable in operation than LobithW and LobithNN.

4 Limitations and recommendations for future works

Based on the experiments and results, limitations of this project are identified, along with recommendations for future research and improvements. On a general note, the quality of LobithW and LobithNN can be considered quite good (the maximum MAE is always below 80m^3/s at +48 hours). Therefore, we advise that future works will focus on aspects such as operational reliability.

4.1 Operational reliability improvements

Implementing the model in an operation system is just a beginning. It is crucial to ensure the deployed model performs as well in an operational setting as it does in a research environment. Broadly speaking, an ML system can encounter three types of issues: input issue, prediction issue, and system issue.

The first and most key step in addressing these issues is monitoring.

- **Input monitoring**. Data skews can lead to performance drops or complete system failures. Data skews occur when the model's training data does not represent the live data. This can happen because of four reasons:
  - The training data is designed incorrectly: distributions of the variables in our training data do not match the distribution of the variables in the live data.
  - A feature used during training is temporarily or permanently unavailable in the operational setting.
  - Research/operation data mismatch: such as using BfG data in model training in research environment and using Matroos data in operation.
  - External data dependencies: for example, changes in upstream water level data from external systems, if not communicated, can affect predictions.

- **Prediction monitoring**. Model performance should be monitored automatically or manually by computing evaluation metrics and/or comparing the model prediction distribution using statistical tests (e.g., mean, median, standard deviation, max/min values). This can be implemented using the RWsOS Verification Dashboard.

- **System monitoring**. This involves monitoring within the realm of software engineering. It is great that RWsOS already has extensive monitoring capabilities 24/7.

Together with the potential issues mentioned above; to enhance operational reliability, we suggest further research on the following aspects:

- **Handling missing input features**. Investigate methods to deal with features that are temporarily or permanently unavailable.
  - For temporary issues, methods such as interpolation or extrapolation can be used, as already implemented in the system. Or re-creating the feature by combining other existing features after studying their relationships.
  - For permanent issues, methods could include replacing the feature with a similar available variable or removing it, both of which may require model retraining.

  - Another proactive approach is training ensemble/backup models, each using a subset of input features. If one feature is missing, the models that do not rely on that feature can still function, providing backup.

- **Model performance drop (staleness)**. Environmental shifts, such as climate change, can alter discharge behaviors at Lobith, causing model performance to degrade over time. Further research is recommended on maintaining model accuracy in production through periodic model retraining and other strategies to adapt to environmental changes.

4.2 Model improvements

1 Improve extreme discharge range.

The best model developed in this project shows significant improvement over LobithW and Old LobithNN for normal discharge ranges. However, for extreme discharge ranges, especially high flows, the model has potential for further enhancement. We recommend the following steps:

- Investigate the low/high flow patterns at Lobith in more detail.
- Improve the model for extreme discharge ranges by:
  - Applying an ensemble model approach with two models having different loss functions, one trained for normal and low flows, and another for high flows.
  - Exploring methods for high flow extrapolation within this model.

2 Input features.

The best model developed in this project, uses the water level at Koblenz station as an input feature. However, internal experts have noted that Koblenz station experiences a backwater issue, which might make its water level measurements less representative of reality. It is suggested to:

- Investigate the backwater issue at Koblenz station.
- Explore the potential of using upstream stations in tributaries as a substitute for Koblenz for peak discharge prediction.

In Test Case 4, we explored adding observed and/or forecasted precipitation from nearby upstream catchments to enhance model performance. This approach was not successful, likely because adding only precipitation does not enable the model to learn hydrological processes effectively and may confuse it. Therefore, it is recommended to explore adding more process information such as evaporation, temperature, other hydrological variables like groundwater levels, and catchment hydrological characteristics, along with precipitation, to help the model learn the hydrological system processes more effectively.

3 Model target.

In this project, we chose to forecast discharge at Lobith, as historical water level measurements were considered less reliable due to changes in the riverbed. However, using upstream water levels as input to predict discharge adds complexity to the relationship between inputs and target. We hypothesize that the model might more easily learn the relationship between upstream water levels and the water level at Lobith. Therefore, we recommend to test and compare the effectiveness of predicting water level versus discharge to determine the most accurate and useful model target, given that the observation data of water level at Lobith is valid and has sufficient long record for model training.

4 Model selection.

This project uses a LSTM-based model architecture based on the previous work of LobithNN. However, from a standard machine learning model development perspective, model selection should be based on 1) defined problem, 2) testing multiple potential model structures, and 3)

balancing model accuracy and interpretability/explainability. We recommend exploring other model structure options, such as a multi-linear regression model (MLR) based on the same setup as the best model we developed in this project.
MLR model may offer better extrapolation and are easier to interpret compared to deep learning models. Additionally, Automated machine learning (AutoML[2]) tools can be used to set baseline for comparison.

5 Model uncertainty.
For the application of machine learning (ML) model in operation, it is valuable to provide a prediction interval that captures the uncertainty of the forecasted values. Future research could explore approaches to quantify ML model uncertainty, such as employing an ensemble of models with varying hyperparameters, applying a Monte Carlo approach, or using conformal prediction technique.

4.3 Broader perspectives

1 Standardize AI model development procedure.
Currently, formal workflows and tools for structuring AI projects within RWsOS are not available. This lack of standardization poses a least two potential problems, such as inability to assess project quality against standardized benchmarks and the time-consuming definition of project workflows at the start of each project. Therefore, we recommend developing standardized workflows and tools for AI projects within RWsOS. This will improve project quality, efficiency, and the ability to supervise these projects effectively.

2 Generalize to other locations.
The knowledge and methods developed in this project could be valuable for other basins besides the Rhine. We recommend developing a generic approach for X-day ahead discharge or water level forecasting that can be applied to other basins.

[2] https://learn.microsoft.com/en-us/azure/machine-learning/concept-automated-ml?view=azureml-api-2

5 Conclusion

This project aims to improve the model for short-term (2 days ahead) discharge forecasting at Lobith with an hourly timescale and improved quality over LobithW. Through experimentation with various test cases, we identified the optimal setup for the new LobithNN model. This model leverages water levels from six reliable upstream stations and historical discharge data at Lobith, trained on BfG data and operationalized with Matroos data. The new LobithNN model exhibits improved performance over LobithW and Old LobithNN in normal discharge ranges across nearly all lead times. Future works are recommended to improve operational reliability, enhance model accuracy, and standardize and generalize machine learning workflows and tools.

A Plan van aanpak doorontwikkeling LobithNN

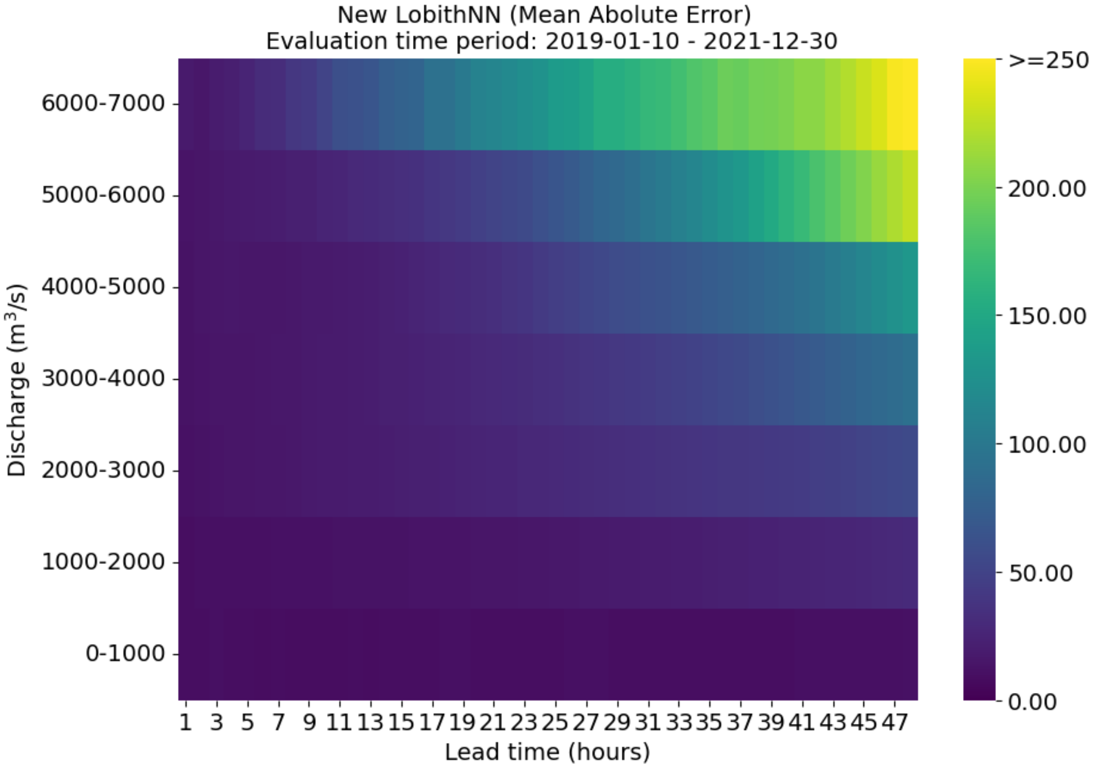
11209230-024-Plan
van aanpak dooront

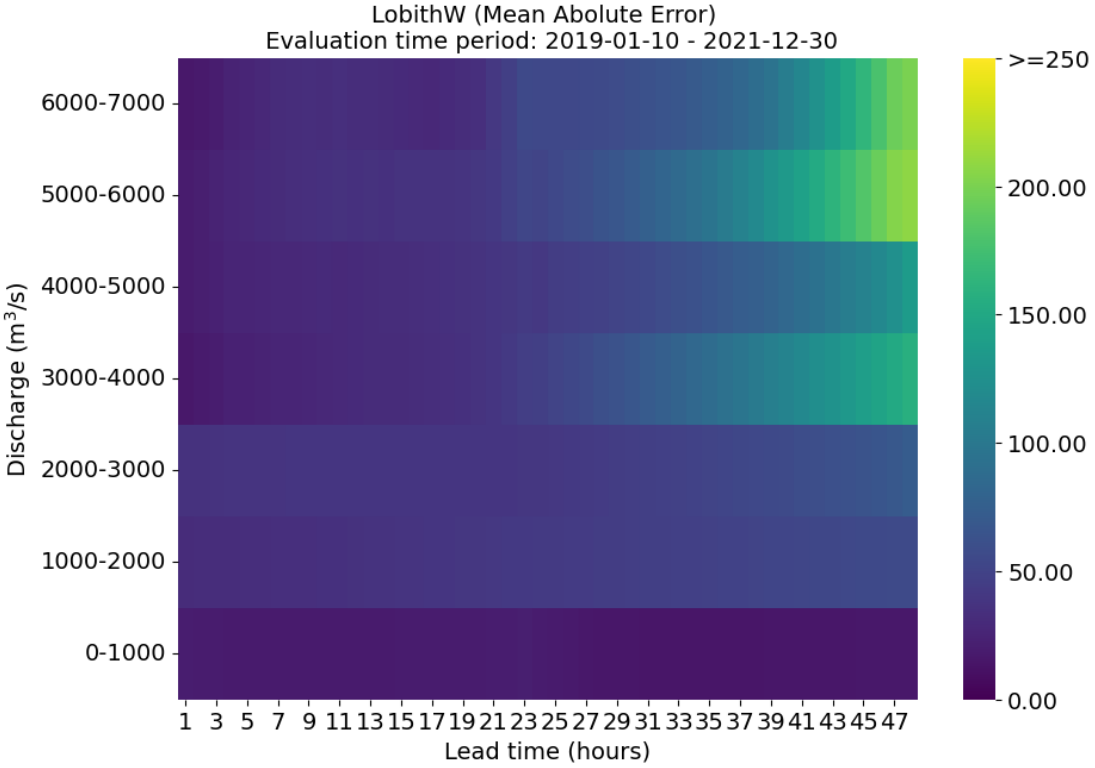
B References

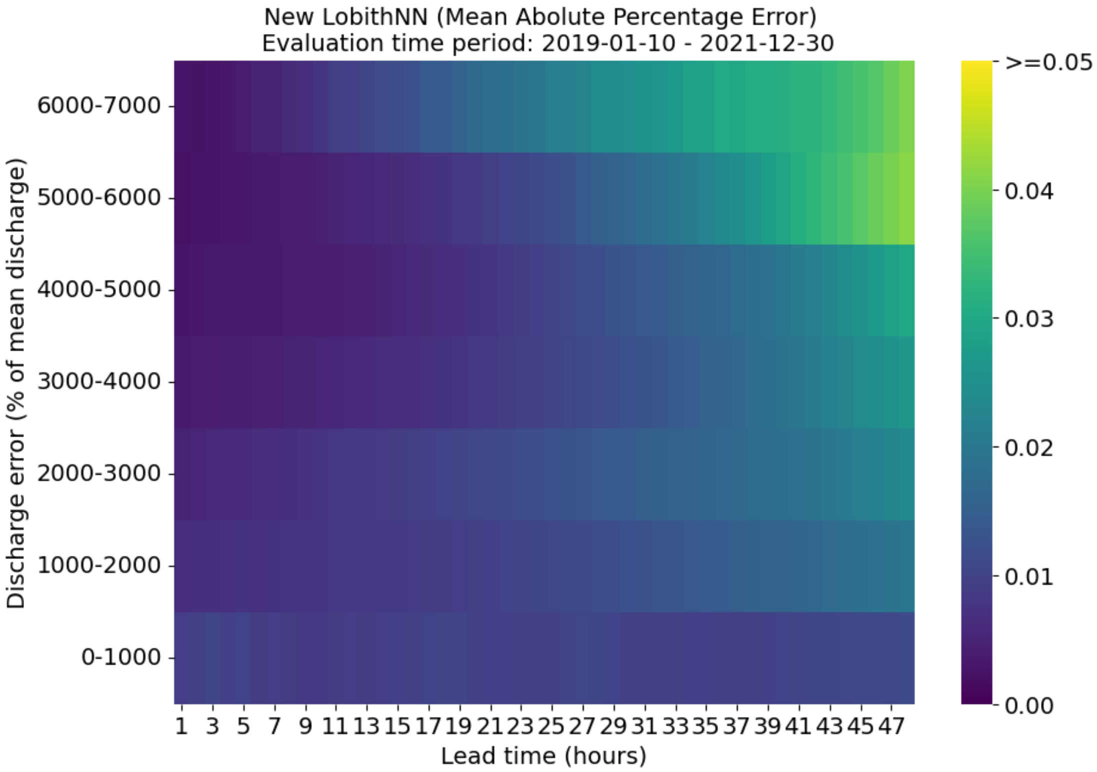
Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." *Neural computation* 9.8 (1997): 1735-1780.

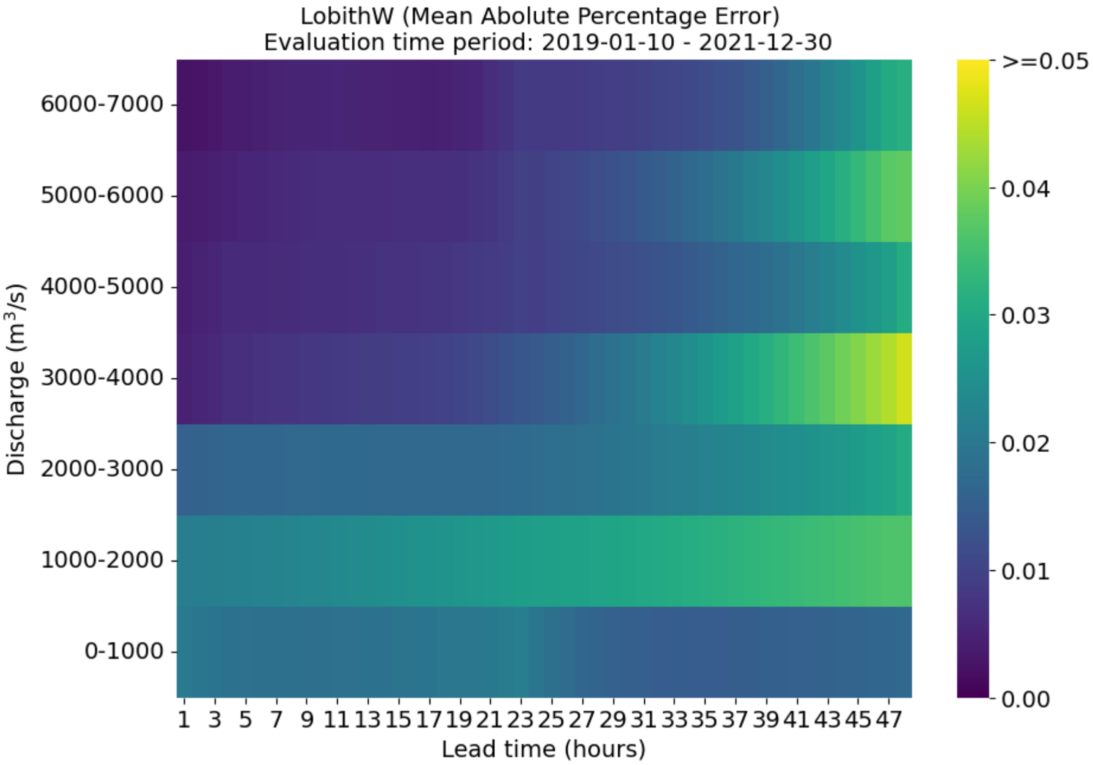
Lundberg, S. M., and Lee, S. I. (2017). A unified approach to interpreting model predictions. Adv. Neural. Inf. Process Syst. 30, 1–5. doi: 10.48550/arXiv.1705.07874

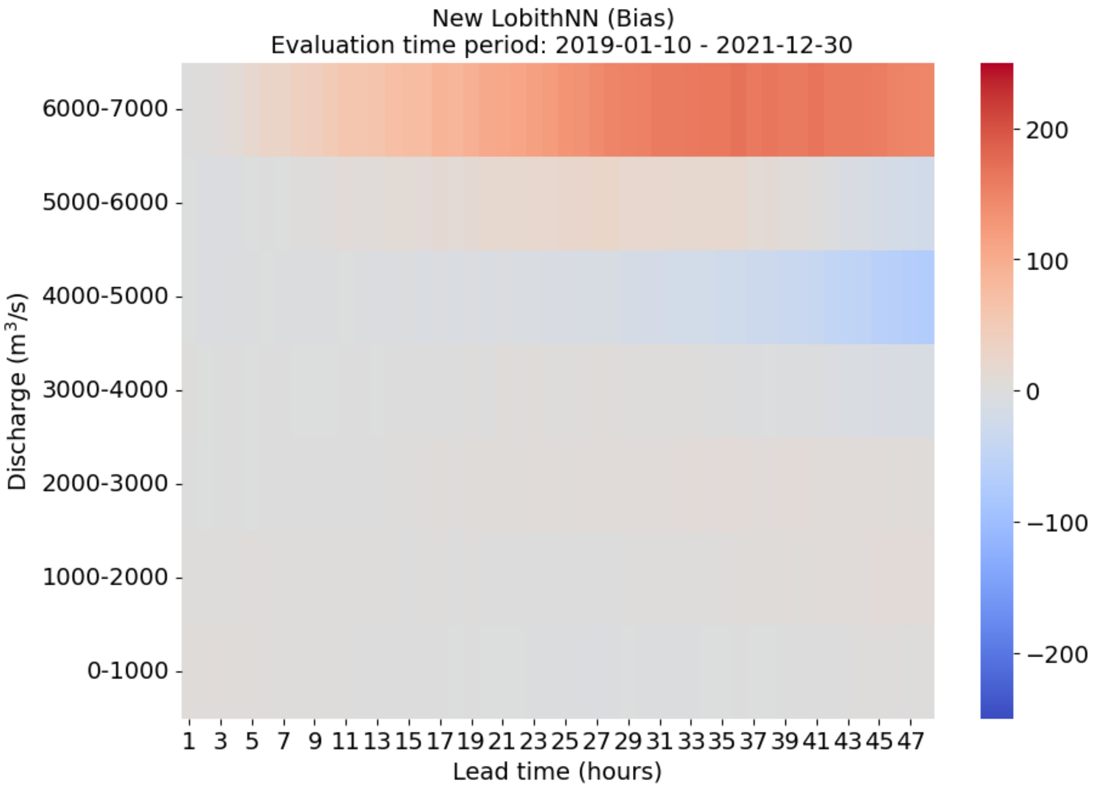
C Additional figures

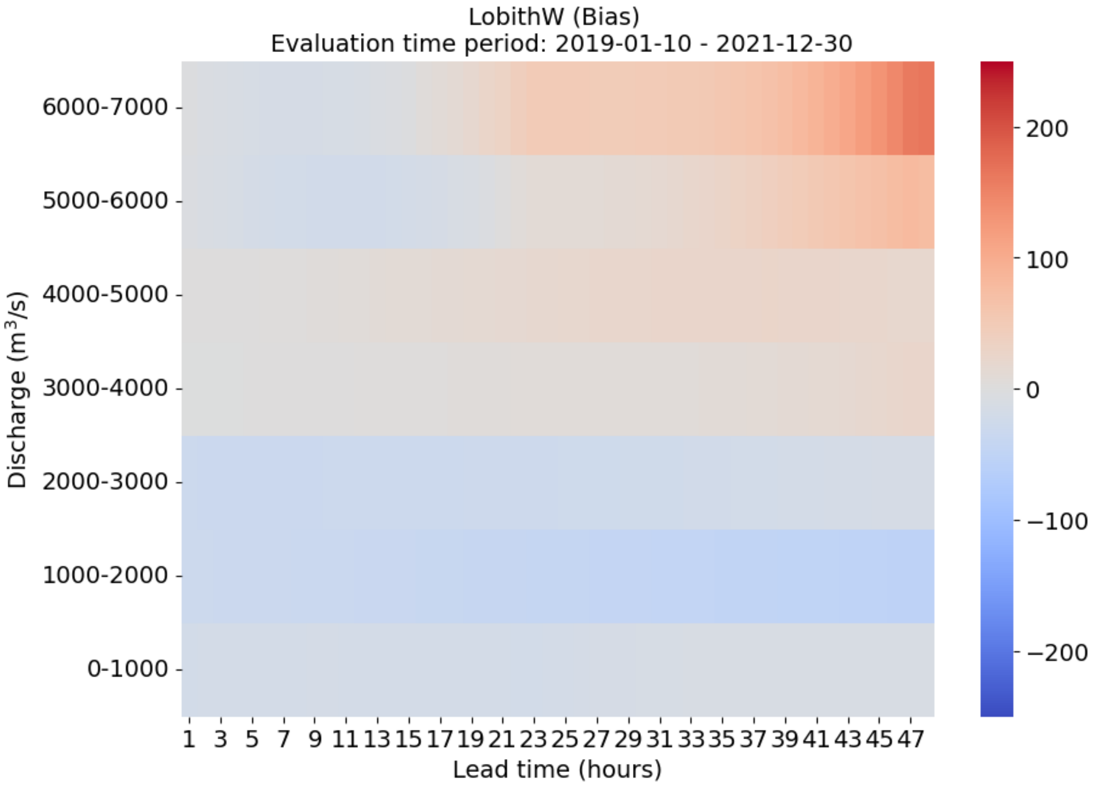


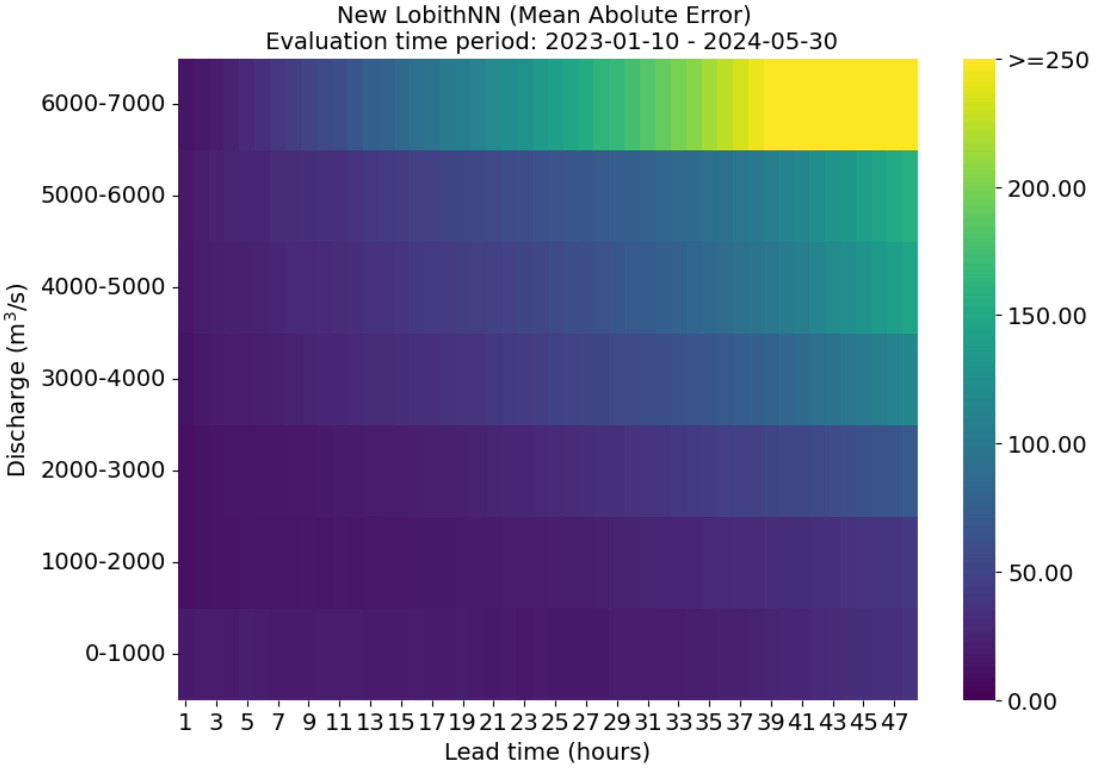


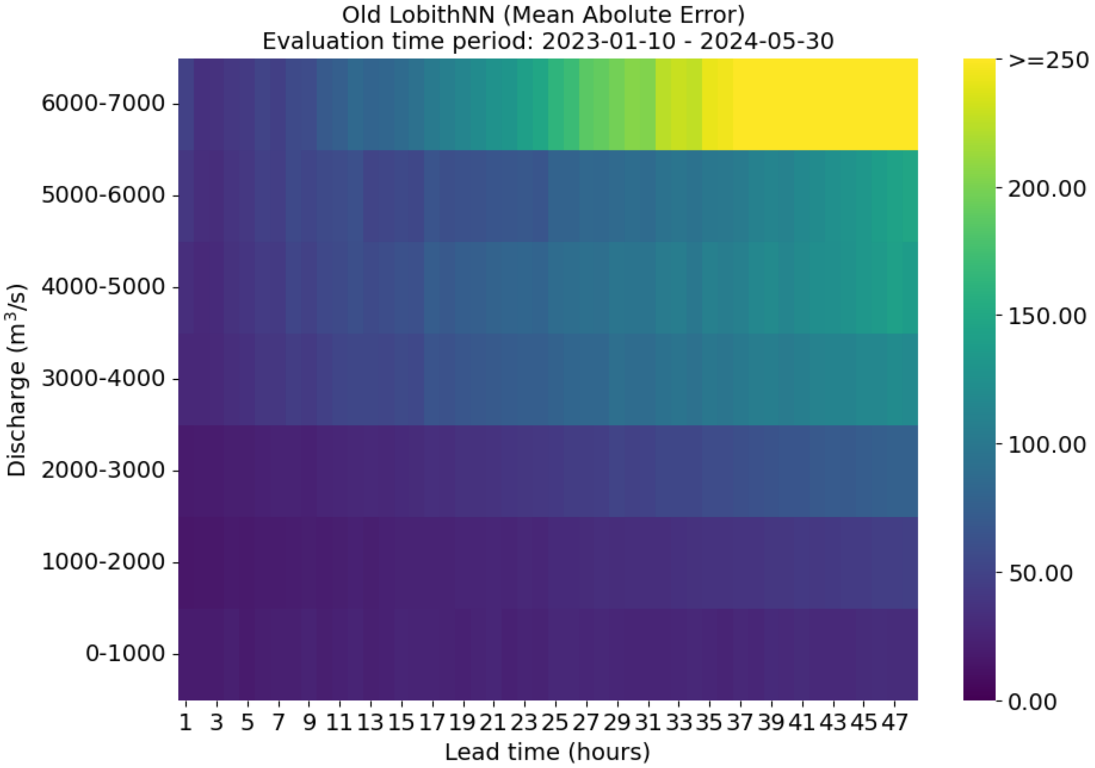

New LobithNN (Mean Abolute Percentage Error)
Evaluation time period: 2019-01-10 - 2021-12-30
Discharge error (% of mean discharge)
6000-7000
5000-6000
4000-5000
3000-4000
2000-3000
1000-2000
0-1000
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47
Lead time (hours)
>=0.05
0.04
0.03
0.02
0.01
0.00

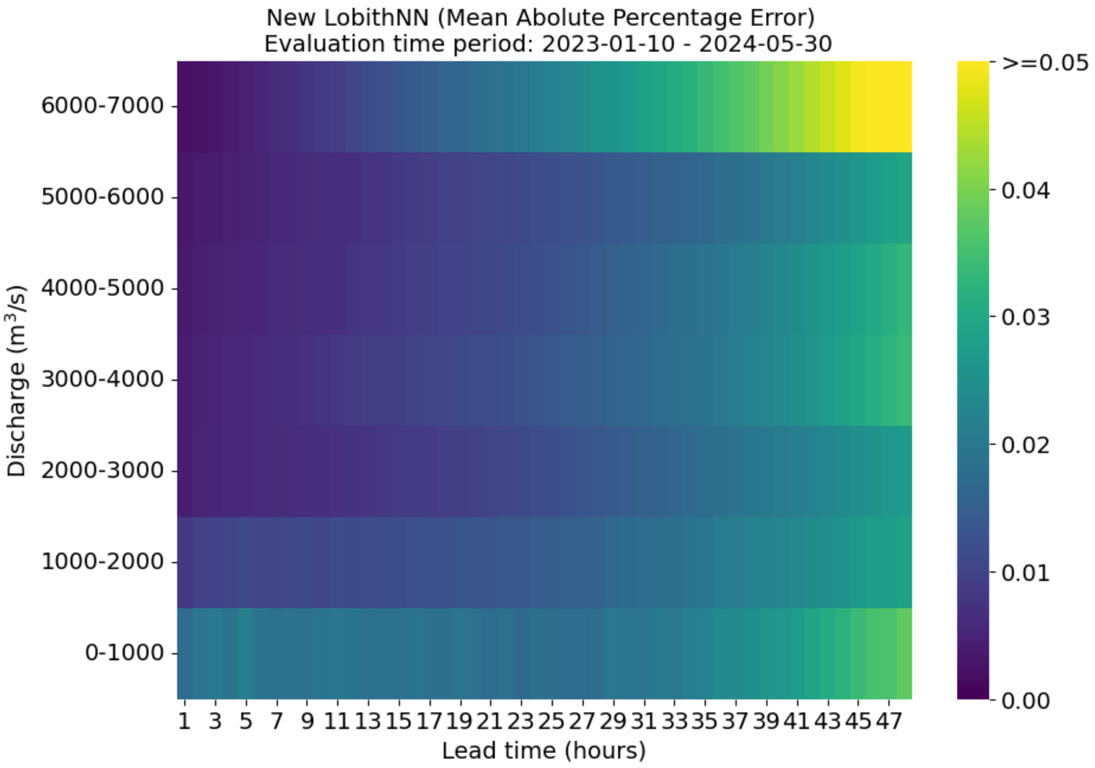

LobithW (Mean Abolute Percentage Error)
Evaluation time period: 2019-01-10 - 2021-12-30
Discharge (m^3/s)
6000-7000
5000-6000
4000-5000
3000-4000
2000-3000
1000-2000
0-1000
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47
Lead time (hours)
>=0.05
0.04
0.03
0.02
0.01
0.00

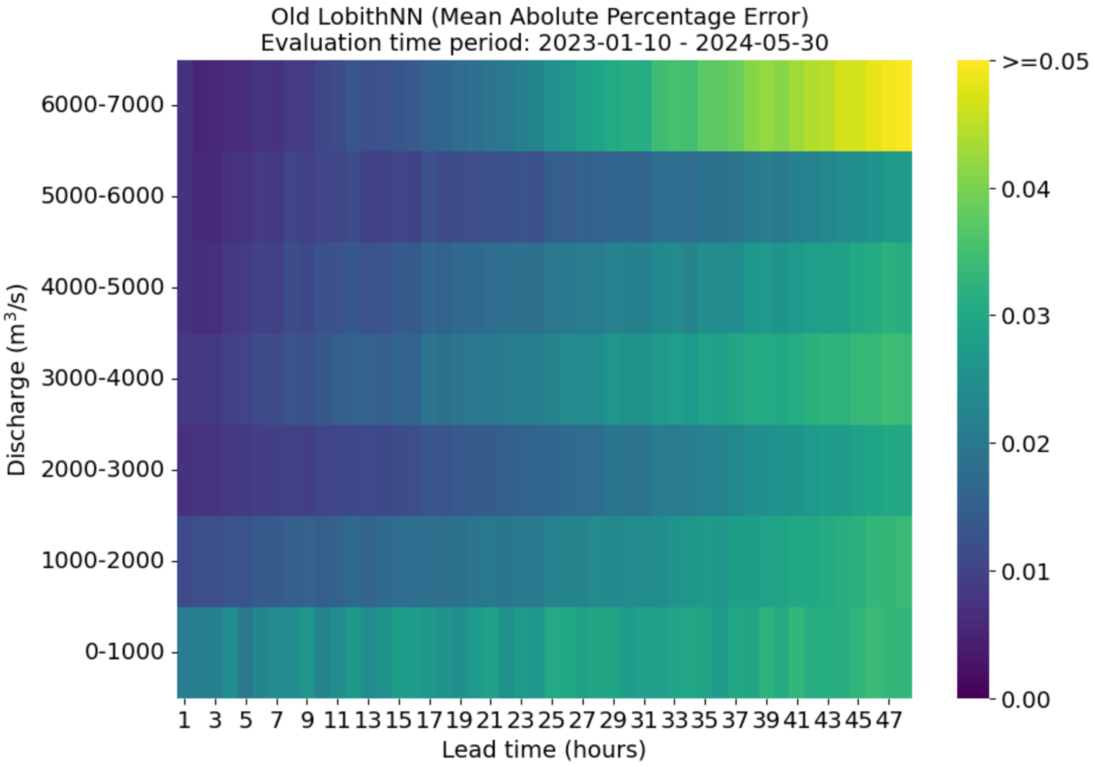

New LobithNN (Bias)
Evaluation time period: 2019-01-10 - 2021-12-30
6000-7000
5000-6000
4000-5000
3000-4000
2000-3000
1000-2000
0-1000
Discharge (m^3/s)
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47
Lead time (hours)
200
100
0
−100
−200

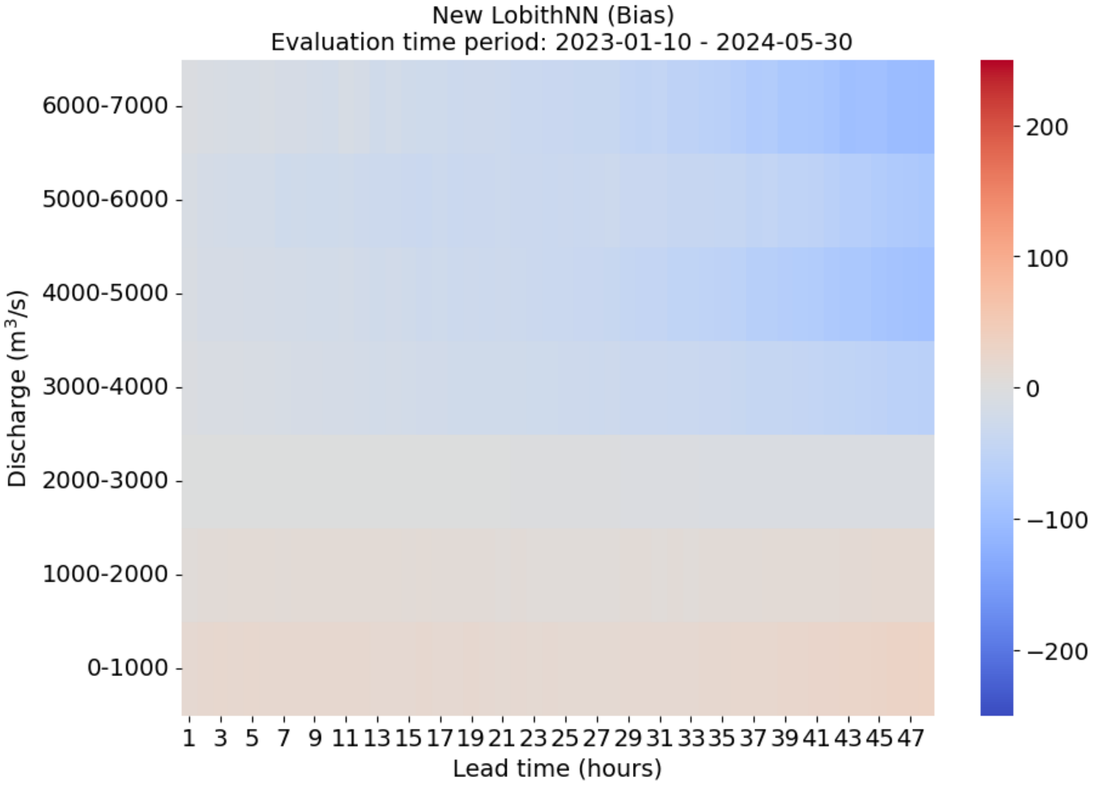

LobithW (Bias)
Evaluation time period: 2019-01-10 - 2021-12-30
6000-7000
5000-6000
4000-5000
3000-4000
2000-3000
1000-2000
0-1000
Discharge (m^3/s)
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47
Lead time (hours)
200
100
0
−100
−200

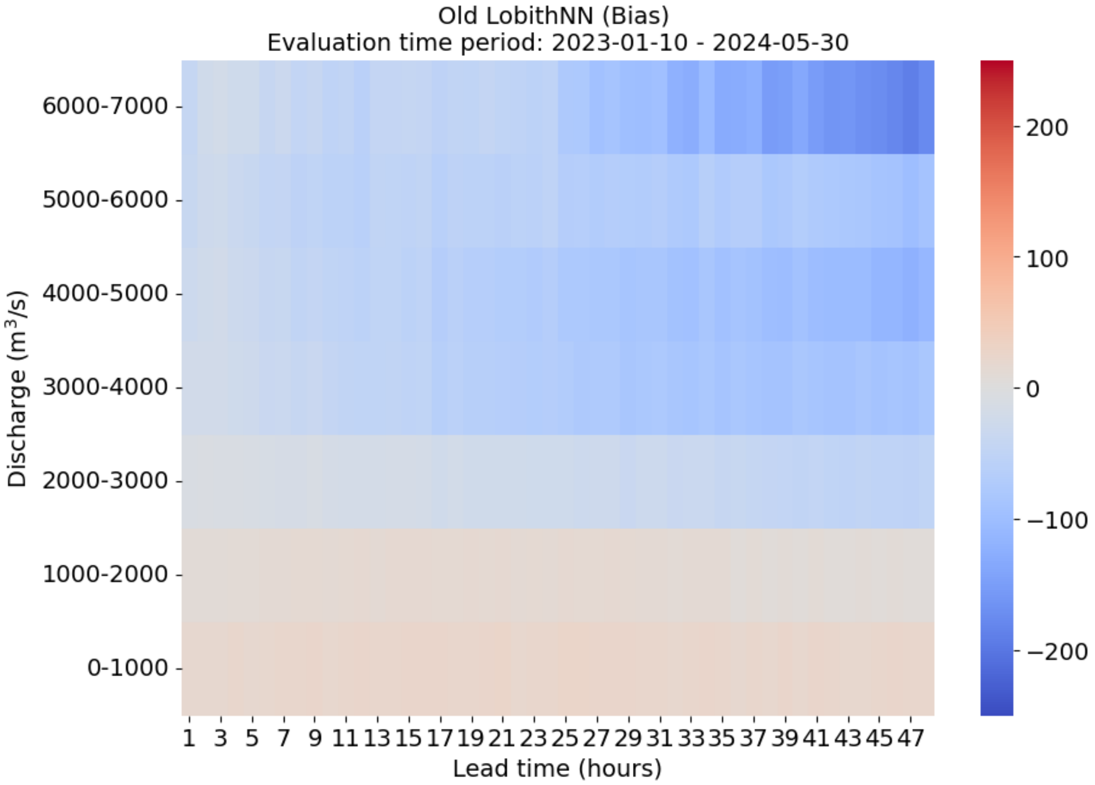

New LobithNN (Mean Abolute Error)
Evaluation time period: 2023-01-10 - 2024-05-30
Discharge (m^3/s)
6000-7000
5000-6000
4000-5000
3000-4000
2000-3000
1000-2000
0-1000
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47
Lead time (hours)
>=250
200.00
150.00
100.00
50.00
0.00


Old LobithNN (Mean Abolute Error)
Evaluation time period: 2023-01-10 - 2024-05-30
Discharge (m^3/s)
6000-7000
5000-6000
4000-5000
3000-4000
2000-3000
1000-2000
0-1000
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47
Lead time (hours)
>=250
200.00
150.00
100.00
50.00
0.00


New LobithNN (Mean Abolute Percentage Error)
Evaluation time period: 2023-01-10 - 2024-05-30
6000-7000
5000-6000
4000-5000
3000-4000
2000-3000
1000-2000
0-1000
Discharge (m^3/s)
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47
Lead time (hours)
>=0.05
0.04
0.03
0.02
0.01
0.00


Old LobithNN (Mean Abolute Percentage Error)
Evaluation time period: 2023-01-10 - 2024-05-30
6000-7000
5000-6000
4000-5000
3000-4000
2000-3000
1000-2000
0-1000
Discharge (m^3/s)
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47
Lead time (hours)
>=0.05
0.04
0.03
0.02
0.01
0.00


New LobithNN (Bias)
Evaluation time period: 2023-01-10 - 2024-05-30
Discharge (m³/s)
6000-7000
5000-6000
4000-5000
3000-4000
2000-3000
1000-2000
0-1000
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47
Lead time (hours)
200
100
0
−100
−200


Old LobithNN (Bias)
Evaluation time period: 2023-01-10 - 2024-05-30
Discharge (m³/s)
6000-7000
5000-6000
4000-5000
3000-4000
2000-3000
1000-2000
0-1000
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47
Lead time (hours)
200
100
0
−100
−200

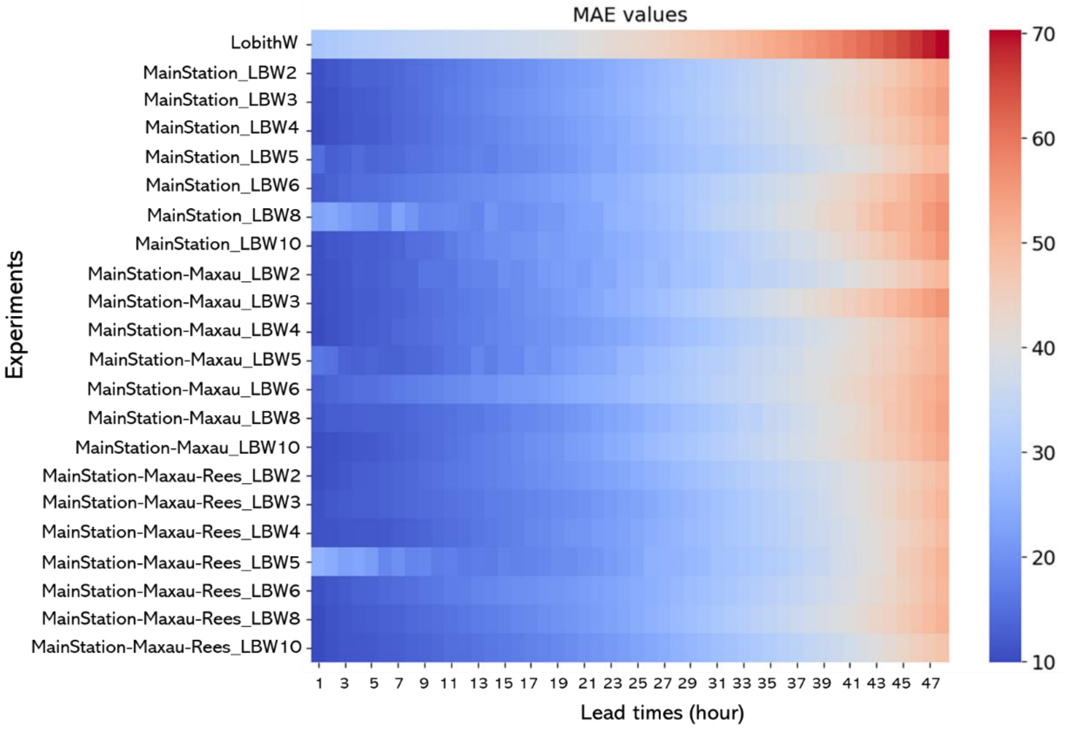


Figure: MAE results of experiments on different look-back windows (LBW) of 2, 3, 4, 5, 6, 8, 10 days with the reduced number of stations. MainStation stands for eight upstream stations: Rees, Wesel, Düsseldorf, Köln, Andernach, Koblenz, Kaub, and Maxau. MainStation-Maxau means all the MainStation except Maxau. MainStation-Maxau-Rees means all the MainStation except Maxau and Rees.

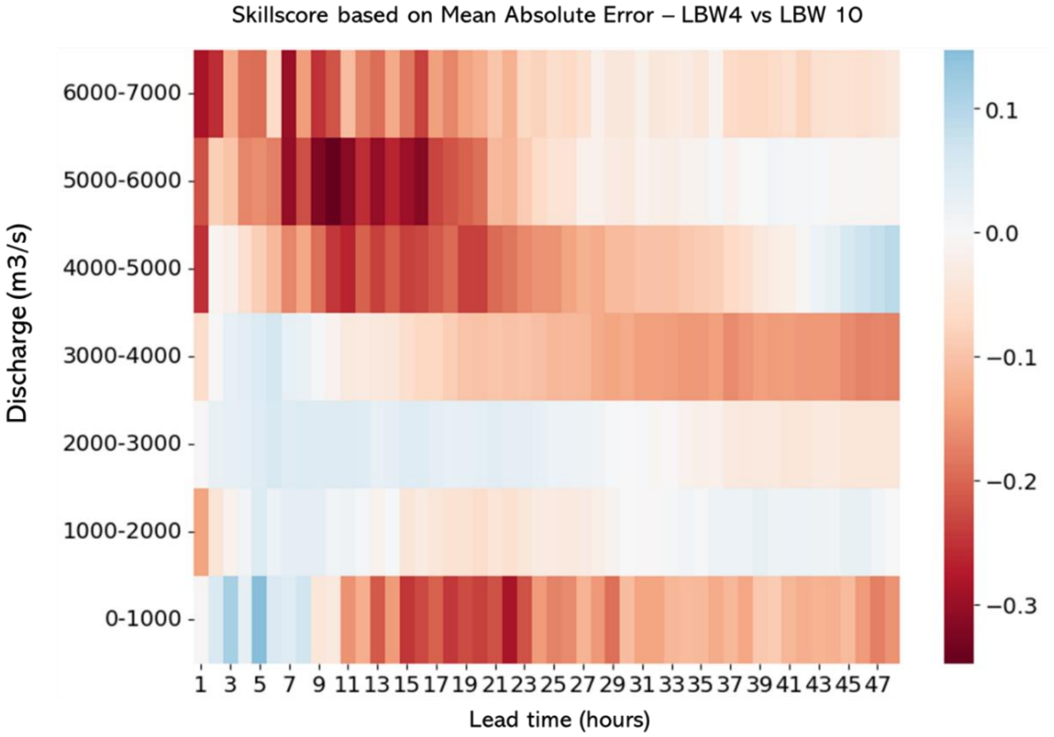


Figure: Heatmap showing the skill scores based on the mean absolute error compares the model setup with six stations using a LBW of 4 days versus the model setup with an LBW of 10 days. Positive values (blues) indicate skill of LBW4 over LBW10. Negative values (reds) indicate skill of LBW10 over LBW4.

D Test case results (more detail)

Testcase 1: less upstream stations
Based on the reliability of BfG data from upstream stations, we selected eight upstream stations: Rees, Wesel, Düsseldorf, Köln, Andernach, Koblenz, Kaub, and Maxau. We further refined the selection based on system understanding and feature importance analysis. For example, Rees and Wesel are close to each other and have highly correlated water levels, and Maxau's water level did not contribute significantly to the results. The experiment results indicate that removing Rees and Maxau from the inputs improves model performance.

As a result of this testcase, we limited the stations to six: Wesel, Düsseldorf, Köln, Andernach, Koblenz, and Kaub.

Additional test case: different look back window
We experimented with look-back windows (LBW) of 2, 3, 4, 5, 6, 8, 10, 15, and 20 days with the reduced number of stations. Among these, LBW of 4 and 10 days resulted in the best performance, with LBW of 10 days showing slightly better performance than LBW of 4 days across most lead times and discharge ranges. See Appendix C for experiment result figures. Although the model setup with an LBW of 4 days involves significantly fewer input time steps compared to the 10-day LBW setup—potentially enhancing input reliability in operational settings—we opted for the LBW of 10 days to ensure more consistent model performance.

Testcase 2: training on time-differenced discharge
Predicting the time-differenced discharge (*dQ*) instead of the discharge (*Q*) directly does not enhance the model performance. This suggests that the model struggles to effectively learn the relationship between the input features (i.e., water levels at upstream stations and historical discharge at Lobith) and the target dQ.

Testcase 4: adding precipitation as input
Adding observed and/or forecasted precipitation from nearby upstream catchments does not enhance the model performance. One probable reason for this, when only adding observed precipitation, is that the precipitation information is already captured in the upstream station water levels. Therefore, adding separate precipitation data does not provide significant added information for the system to predict discharge at Lobith. Another potential reason is that simply adding precipitation does not allow the model to truly learn how the hydrological system works. In hydrological modelling, besides precipitation, factors such as potential evaporation, temperature, and system storage are crucial for understanding the system. Without incorporating these processes into the ML model, it cannot learn the system effectively. Instead, the addition of only precipitation might confuse the model.

Attachment(s)
n:\Projects\11209000\11209230\C. Report - advise\024-MO-06 Dooroontwikkeling hydrologisch modelinstrumentarium\

Deltares is an independent institute for applied research in the field of water and subsurface. Throughout the world, we work on smart solutions for people, environment and society.

Deltares

www.deltares.nl