Prediction of Discharges from Polders to ‘Boezem’ Canals with a Random Forest and an LSTM Model

Improving Inputs of the Decision Support System of the Hoogheemraadschap van Delfland

Josine van Marrewijk

Delft University of Technology

Prediction of Discharges from Polders to ‘Boezem’ Canals with a Random Forest and an LSTM Model

by

Josine van Marrewijk
(4675878)

To obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on November 2nd at 11:00 AM.

Dr. Riccardo Taormina	Delft University of Technology
Dr. Marc Schleiss	Delft University of Technology
Prof.dr.ir. Remko Uijlenhoet	Delft University of Technology
Ir. Joost Driebergen	HKV
Ir. Sjoerd Gnodde	Hoogheemraadschap Delfland
Version	Thursday 26th October, 2023
Faculty:	Civil Engineering and Geosciences, Delft

Cover: Boezem canal ‘Groeneveldsche- of Monsterwatering’, September 26th, 2023.

Preface

The past eight months have been the most interesting and challenging part of my study. I have learned new things, for example making machine learning models, which I would not believe if you told me a few years ago. I am grateful for the experience working at a company as HKV with a lot of experienced colleagues and learning about the operational water management of the area where I grew up.

I would like to take this opportunity to thank my committee. First of all, Riccardo, his enthusiasm about AI and the case study motivated me throughout the whole process. Riccardo gave me confidence that I was able to do this and that what I was doing was relevant. Then Marc and Remko, who contributed to my thesis by providing constructive feedback during the process and on the end-result. I would also like to thank HKV and especially Joost, Ron and Thomas for their supervision and their insightful perspectives. Next, I would like to thank everyone at the Water Authority of Delfland who took the time to help me understanding the challenge of the operational water management in the area. I very much enjoyed talking to different people of Delfland about the possibilities of data-driven models in their organization and the considerations that need to be taken into account when constructing a model. In particular I would like to thank Sjoerd, who was always ready to help me out whenever I had a question or needed to discuss some results or doubts. His engagement, encouragement, and enthusiasm helped me greatly throughout the process.

Working on my thesis was a first introduction to the 'working' life and was better than I expected. I am going to miss the coffee breaks and discussions with my fellow watermanagement students at the 'hok'. I have also realized that even though you are working on your own project, it felt like something I was doing together with my friends, just like the time we were studying together during the bachelor. The end of the thesis is a relief, but at the same time a realization that this unforgettable and amazing time is coming to an end.

Last but not least, I would also like to thank my family for their support and love. A special thanks to my roommates, the 'Molsmeiden', who have made my student life complete. You feel like my second family, cheering me up, slowing me down and motivating me, for more than 5 years already.

To conclude, I would like to take this opportunity to express my gratitude for all the valuable experiences of the past 6 years. For example, meeting new people during my exchange in Milan, teaching mathematics at a high school and learning about the challenges of installing new water supply systems in Uganda during a Multi Disciplinary Project. I hope I can continue in this field and contribute to a sustainable future, in which live without water abundance or shortage, everywhere in the world.

Josine van Marrewijk
Delft, October 2023

Abstract

In this research the possibilities of the application of machine learning models at a water authority in the Netherlands are studied. This study was performed at the 'Hoogheemraadschap van Delfland', which is the organization responsible for the operational water management of the area between Rotterdam, The Hague and Zoetermeer.

One of the tasks of the water authority is to maintain the water level in the boezem canals close to the level of -0.43 m NAP. This is done automatically with a Decision Supportive System (DSS) which operates the boezem pumps. An important input for the DSS are the estimated discharges from the polders. Currently, these discharges are predicted with a Sobek RR model. This is a combined conceptual and physical model that has been lumped because of the long computational times. The simplification of this model causes a less accurate prediction with a lower spatial resolution.

The possibilities of machine learning models for the prediction of discharges from the polders are researched by evaluating the performance of two models. Using these models, the sum of the discharge in the next 2, 8 and 12 hours is predicted. First, a random forest (RF) regression model is evaluated, followed by the evaluation of a long short-term memory (LSTM) neural network for the prediction of the 12 hourly sum of the discharge. The machine learning models for the pumping stations should be optimized and trained based on the specific data. Another requirement for the model to perform well is that the data should be complete and there should be no major changes in the system.

This research has shown the potential of machine learning models for the prediction of discharge for the considered pumping stations in the case area consisting of the Duifpolder, Holierhoekse- and Zouteveensepolder and the Vlaardingen-Holierhoek polder. This case area is clustered in the Sobek RR model as node 49. The RF and LSTM model are compared to the current Sobek RR model, the machine learning model of Delfland (ReRengAI) and a naïve model by calculating the root mean squared error (RMSE) for the last year of the dataset. For the prediction of the 2 hourly sum of Node 49 the RF model performs the best. Additionally, the performance of the RF model for the 12 hourly sum is satisfactory with a RMSE of 11,071 m^3, though using a deep learning model (LSTM) the performance improved to a value of 10,181 m^3 for the RMSE. With both these machine learning models it is thus possible to estimate the discharges, and we could apply them also in polders in which the data on the discharges, water levels and precipitation is available.

Machine Learning models are known as black-box models and are hard to explain or interpret. The consequences of these characteristics is that the practical implementation of these new models, despite good model results, is challenging. Technical recommendations for implementation ML models are improving the quality and availability of the data, increasing the interpretability and explainability of the model, combining multiple objectives in the new model or combining a ML model with a physical model. Organizational recommendations are improving the knowledge about these models within the organization, studying the advantages of these models in comparison to the current model and involving different departments of the water authority in the development of these new models. This study found that there is potential for the development of ML models since the used models show good results for the discharge prediction. On top of that, the water authority is interested in the ways how the water system can become more 'future-proof'. An example of how a water system can become more future proof is by optimizing the pumping operation based on electricity prices, water quality and ecology. If next to the improved estimating of the discharges, also other purposes are implemented in the new machine learning models, the urge and the willingness for the development of these models at water authorities will increase and then be accelerated.

Contents

List of Figures

List of Tables

Nomenclature

Abbreviations

Abbreviation	Definition
AI	Artificial intelligence
DL	Deep learning
DUI	Duifpolder
FEWS	Flood Early Warning System
HZP	Holierhoekse- en Zouteveensepolder
IQR	Interquartile range
LIME	Local Interpretable Model-agnostic Explanations
MAE	Mean Absolute Error
ML	Machine Learning
MSE	Mean Squared Error
NAP	Normaal Amsterdams Peil (‘Normal Amsterdam Level’)
NN	Neural Network
NSE	Nash–Sutcliffe efficiency
LGB	Light Gradient Boosting
PINN	Physics Informed Neural Network
RF	Random Forest
RMSE	Root mean squared error
RTC	Real Time Control
SHAP	SHapley Additive exPlanations
VHK	Vlaardingen-Holierhoek polder

Definitions

Machine learning terminologies

Term	Definition
Activation function	Produces the ranges in which the output is scaled. The function should be differentiable in order to employ backpropagation.
Algorithm	A set of steps that a machine learning model follows to learn from data and make predictions.
Backpropagation	Gradient descent on individual errors. The predictions of the neural network are compared with the desired output and then the gradient of the errors with respect to the weights of the neural network is computed. This provides a direction in the parameter weight space in which the error will become smaller (Rumelhart, 1986).
Epoch	An iteration to minimize the loss function (Razavi, 2021).
Feature	An input variable that presents a aspect of the data.
Gradient descent	An optimization algorithm used to minimize the loss function by adjusting the model’s parameters in the direction of steepest descent.
Hyperparameters	These are the parameters that determine the learning process and the model’s architectures.
Label	The target variable, which is the output variable of the model.

Term	Definition
Pruning	The termination of unpromising trails (automatically early stopping) (Akiba et al., n.d.).
Regularization	Prevents overfitting and reduce complexity by adding a penalty to the network (example: early stopping and dropout) (Shen, 2018).

Dutch terminologies

Beslissing ondersteunend Systeem (BOS) = Decision-Supporting System (DDS).
Boezemgemaal = Pumping station that pumps water outside of the system into the sea or a river.
Boezemkanaal = Drainage canal where the water level is being controlled.
Gemaal = Pumping station.
Inslagpeil = Pump initiation stage (water level at which the pump is switched on).
Peilbesluit = Water level ordinance.
Poldergemaal = Pumping station that pumps water into the ’boezem’ canals.

1 Introduction

1.1. Research motivation

One of the most exciting technologies of this time is Artificial Intelligence (AI). It has the ability to process data and extract useful information from it (Shen, 2018). This allows us to create algorithms that can make predictions and find relations between features. In water management, AI can be used for applications such as finding leaks, monitoring water quality, and predicting discharges. Machine Learning (ML) models, a type of AI, can be used for simulating rainfall-runoff because they can handle large amounts of data and make accurate predictions for new situations. The water authorities in the Netherlands collect a lot of data, which is crucial for a ML model. In the past years, the water authorities are looking into ways to use their data and make their operational systems more ’future-proof’. In this research the possibilities of Machine Learning models for the prediction of the discharges in the operational water management of a polder system are investigated, technically and partly socially.

In this research the considered study area is the Water Authority of Delfland, located in the western part of the Netherlands. One of the objectives of Delfland is to maintain the water level close to the target water level of -0.43 m NAP in the ’boezem’ canals. This is important to avoid flooding of the polder, to assure the navigability of the canals and to make sure the dikes are in good condition. The initially allowed deviation of this water level is 3 cm, meaning that there is a limited buffer capacity in the area. In extreme situations, the water level in the boezem canals can rise up to the alarm value of -0.35 m NAP. In this case the locks in the city centre of Delft need to close.

The water level is being controlled by a Decision Supportive System (DSS), in Dutch referred to as BOS (’Beslissings Ondersteundend Systeem’). This system operates the pumps in case of a ‘normal situation’. The DSS runs different tasks, such as data importation and pre-processing. The DSS is directed by a Flood forecasting and Early Warning System, FEWS, a software suite from Deltares (“About Delft-FEWS - Delft-FEWS - oss.deltares.nl”, n.d.). More information on DSS can be found in Chapter 2. The polder pumps discharge excess water from the polders into the boezem. How much water is coming from the polder is a very important input for the operational control of the ’boezem’ pumps. These discharges are calculated with the Sobek Rainfall Runoff (RR) model. One of the problems is that it takes too long to make a prediction of the discharge for each pumping station separately to use the information as an input for the model. Delfland resolved this by lumping the model. Different polder pumps are clustered and schematized as 1 pump that discharges water at one location. The 138 pumping stations are simplified into 25 pumps to make it possible to calculate the discharges ’faster’. Another problem of the current model is that the measurements of the current state is not taken into account. This means that very useful information that is being monitored, is not used for the prediction of the discharges, which lowers the performance of the model. Lastly, the spatial resolution of the Sobek RR model is lower than in reality, which means that information on at which locations the water is pumped into the system is not taken into account.

The prediction of the amount of discharges from the polders could be improved by calculating the discharge for each polder pump separately. In order to do this, a different type of model needs to be considered. The three commonly used models for simulating rainfall-runoff responses are conceptual models, physically based models, and fully data-driven models (Kratzert et al., 2018). Since there is a lot of data available and the computational time should be decreased, it is suggested to look into data-driven models for predicting the discharge.

The popularity of using data-driven models for time series forecasting, in particular deep learning (DL) models, has increased over the last few years since neural networks (NN) are able to learn dependencies in time series. With deep learning it is possible to find new features and increase the efficiency of the modelling process (Lara-Benítez et al., 2021 Shen, 2018).

Choosing the best type of deep neural network and architecture requires expert knowledge. In the research done by Lara-Benítez et al. (2021) different NN are evaluated for different types of forecasting problems, with LSTM and CNN as the best models. In the research of Kratzert et al. (2018), it is stated that there is potential for using LSTM models to 'learn' dependencies between input and output. This is a specific type of a recurrent neural network (RNN): long short term memory (LSTM) neural network, suggested by Hochreiter and Schmidhuber (1997). This type of neural network is good in processing and predicting sequential data and able to predict multiple outputs. In contrast to the classic RNN, a LSTM model is able to learn long-term dependencies and is therefore suitable to use in rainfall-runoff predictions (Yu et al., 2019). These models were able to capture storage effects in for example catchments with part of the precipitation falling as snow (Kratzert et al., 2018) and showed promising results in ungauged basins (Wilbrand et al., 2023). More details about LSTM models are presented in Chapter 2.5.

It is difficult to obtain insight into the decision-making process of a deep learning model, so also another type of machine learning model is considered. An option is to use a random forest (RF) regression (Breiman, 2001), which is known to be simple, accurate and interpretable. An RF model is commonly used in rainfall-runoff modelling (M. Li et al., 2020). This model is a tree based algorithm that can provide insights in the features that have the highest correlation with the output variable (Qiao et al., 2023). The prediction is made by combining multiple individual decision trees, which makes it a robust model. The RF model is able to deal with complex and large data sets (Muñoz et al., n.d.), which is required to handle the big amount of data that is available. More details about RF models is discussed in Chapter 2.4.

At the water authority of Delfland, there have been some developments in using machine learning (ML) to predict the discharge from the polder pumps. The type of ML that is used is 'light gradient boosting' (LGB) model, which is a type of a supervised machine learning model that uses decision trees. More detailed information on LGB can be found in Appendix A.2. The Hoogheemraadschap of Delfland was positive about the results of this LGB model, but the model is not yet in operational use.

1.2. Problem statement

Currently the amount of water that is pumped from the polders to the boezem is predicted with a Sobek RR model. This model is a combined conceptual/physical model, in which the 128 polder pumps are simplified into 25 clusters. This estimation of discharges from the polders is too course. It happens that a whole cluster is predicted to be discharging to the canals and the 'boezem' pumps immediately react. In reality the discharge is smaller, since not all the pumps are switched on, resulting in an overestimation of the discharge. The result is the activation of the boezem pumps, while there was no need yet. On top of that, the information about the locations of the different pumps is not considered because of this clustering. Another major problem is that the system is not always fully trusted by the operators. The operators tend to check the system regularly, even during the night. This means that even though the operation of the boezem pumps is done automatically, still people are checking the system because they are not sure about the performance of the model.

1.3. Research objective

The goal of this research is to evaluate if it is possible to use machine learning to predict the discharges from the polder pumps and using this predictions for the control of the water level in the boezem canals of ‘Hoogheemraadschap van Delfland’. Managing the water level in the boezem canals is important to avoid flooding of the lower lying polders, to keep the dikes saturated to maintain the strength and make sure the canals are accessible. The operational water management of Delfland aims to keep the water level in the boezem within the margin of -0.46 and -0.40 m NAP. A better prediction of the discharges from the polders is favorable because this improves the input of the model so that the water level can be kept at a more stable level. This is particularly crucial because of the system’s limited buffer capacity. The objective is to improve the prediction of the incoming discharges from the polders. This will be done for each pumping station separately, which will possibly increase the accuracy of the prediction and control of the water level in the boezem canals, since the input (discharge from the polders) will be more precise and it will be possible to take into account the specific location of the inflow of the water. The focus in this research will be mainly on the ‘Duifpolder’, since the water levels and discharges in this polders have been measured consistently. To make it possible to compare the model results to the current Sobek RR model, also the pumping stations in the ‘Vlaardingen-Holierhoek polder’ and the ‘Holierhoekse- en Zouteveensepolder’ are considered for the prediction of the discharges. The four pumps in these three polders are lumped together in the node 49 (as one cluster of the current model). Next to that, some research will be done into the implementation of machine learning models at a water authority, which will provide information on the requirements for a model and give some insight in the attitude towards ML models in general.

1.4. Research questions

The main question of this research is: ‘What is the performance and the added value of using machine learning for the prediction of discharges from the polders to the boezem in Delfland?’.

To evaluate the possibilities, there are three main tasks in this research. First, define an output variable and study the most important features for the prediction of the discharge. Secondly, once the model works, it is important to compare the model with the current models that are used by Delfland in order to find out what the added value of the new model is. The last part of the research focuses on the steps that need to be taken in order to implement a deep learning model.

Subquestions

- SQ1: Which factors need to be considered for modeling the discharge from the polders and which time horizon is relevant?
- SQ2: How can the inlets in the system be estimated and how are the different polders connected?
- SQ3: Which features are important for the prediction of the discharges?
- SQ4: What are the optimal hyper parameters of a random forest and a LSTM model?
- SQ5: What is the performance of a RF and LSTM model compared to the current Sobek RR model and the ReRengAI model?
- SQ6: Which steps must be taken to implement a ML model for the operational control of the water in Delfland?

1.5. Reading guide

For reading this thesis, some background information is useful to better understand the current operational water management of the Water Authority of Delfland. Next to that, it is recommended to read chapter 2.4 and 2.5 if you are new to machine learning and neural networks. The methodology (Chapter 3) consists of four parts. The first part is about data preparation and preprocessing. The second part contains information about the variable that will be predicted with the model and the set-up of the random forest model and the LSTM model. Then the methodology for the optimization and evaluation of the different models is discussed. The last part is about the approach to research the attitude towards machine learning models for the operational control of water in Delfland. In the results, the findings of the research are presented and discussed and the different subquestions are answered. For the reader interested in the implementation of ML and DL models in an organization, the Chapters 3.5, 4.6 and Appendix A.6 are recommended. In Chapter 5 a summary of the limitations and recommendations is given. In the final chapter, the conclusion, the results of using machine learning for the prediction of the discharges and the considerations for implementing these models in Delfland are presented.

2 Background

In this chapter some background information about the study area and the operational water management in Delfland is provided. Next to that, different hydrological models are discussed and in particular, some details about data-driven models like random forest Regression and LSTM models are given. For the reader new to machine learning and deep learning techniques, it is recommended to read Appendix A.2.

2.1. Study area

The water authority (or water board) of Delfland, in Dutch “Hoogheemraadschap van Delfland”, is located in the western part of The Netherlands. The tasks of the water authority are to protect the area against flooding and to ensure the water supply. Next to quantitatively managing the water, another important task of the water board is to monitor and control the water quality and to take care of the waste and water treatment. The waterboard of Delfland, referred to as ’Delfland’, controls the water level in the polders and the so-called ’boezem’ canals. The boezem canals are the waterways like ‘De Schie’, ‘Vlaardingervaart’ and ‘Zuidvliet’ that are used for transportation and as water storage. The target water level of -0.43 m NAP (see Figure 2.1) is controlled in order to avoid flooding of the polder area, to maintain the strength of the soil of the dikes and to assure the navigability.

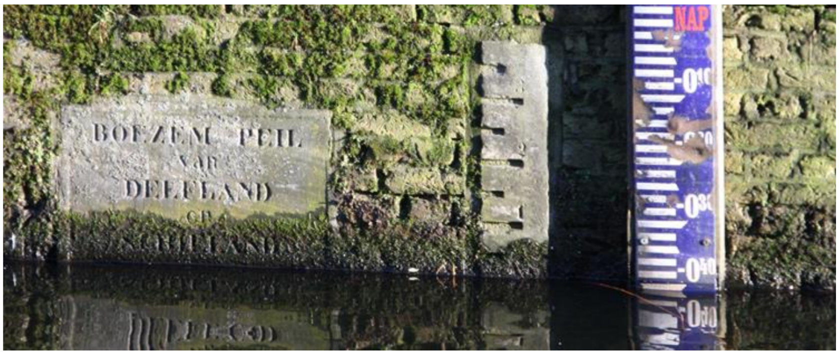


Figure 2.1: Water level in the city center of Delft (Hoogheemraadschap Delfland, 2017)

The area of Delfland consists of several polders and ’boezemland’ these are the higher elevated areas in the western part of the area. Delfland has various landuses and does not have a lot of buffering capacity, which means the response to a dry or wet period can be very quick and maintaining the water level at a constant level can be a challenge.

2.2. Operational water management in Delfland

2.2.1. Water level polder canals

From the polders of Delfland, the excess of water is pumped into the higher water level canals, referred to as 'boezem' canals. The pump are activated automatically based on a certain water level, the so-called 'pump initiation stage' (Dutch: 'inslagpeil'). This initiation stage depends on the location of the polder and can vary over time. For example, in summer time the initiation stage can be slightly higher, to assure al to ensure an adequate groundwater level.

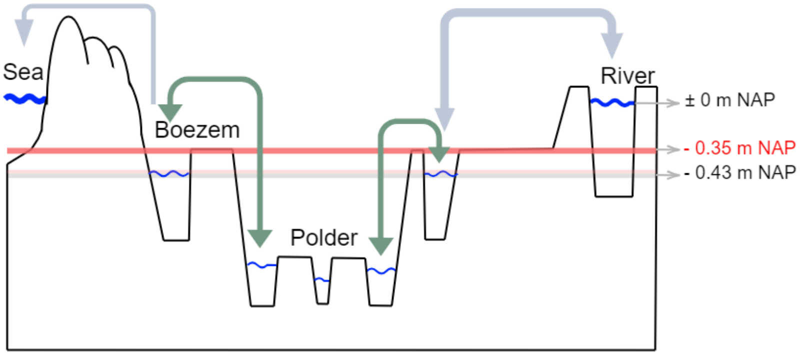
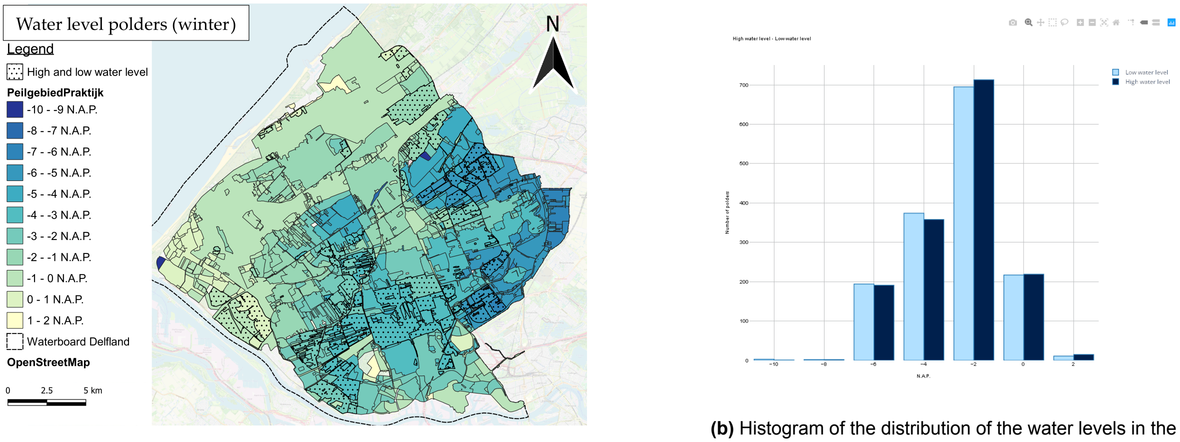


Figure 2.2: Polder system with the threshold water levels.

The water level in the polder canals changes in every polder and also over time. In Figure 2.3a a map with the water levels is presented and in Figure 2.3b the difference of the distribution between the high and low water level is presented in a histogram. In almost 90% of the polders, there is a fixed water level through the year.



(a) Low water level (winter) in the different polders (Delfland, 2022)

(b) Histogram of the distribution of the water levels in the polders during summer (high) and winter (low)

Figure 2.3: Water level in the polders of Delfland (Delfland, 2022)

2.2.2. Water level boezem canals

The objective of the water authority is to keep the water level in the boezem canals as constant as possible, with a maximum deviation of 3 centimeter from the target level of -0.43 m NAP (see Figure A.1.2). If the water level increases to the alarm value of -0.35 m NAP, the sluices in the city center of Delft must be closed to prevent flooding of the area. There are 8 boezem pumps that pump the water out of the boezem into higher located canals, rivers or into the sea, for example into the 'Nieuwe Waterweg', 'Nieuwe Maas' or into one of the neighboring areas. In order to have water of good quality in the boezem and canals, water is let into the polder with these boezemgemalen and the 2 'inlaat'-gemalen, that can pump water into the polder ('inlaat'). There are also 3 pumps that transit the water through the system ('doorvoergemaal'). To maintain the water quality in the boezem canals, water from outside Delfland is let into the system. This is done at several points in the system, such as the boezemgemaal of Winsemius and Dolk, which are further quantified in Appendix A.1.2. The amount of water from the boezem that is let into the polder, is not measured, and therefore difficult to predict. Next to that, there is also water exiting from and going into nearby polders, referred to as seepage. The inlets result in the circulation of water. Stagnant water will result in a poor water quality. This is because of the depletion of oxygen which is essential for the flora and fauna in the water.

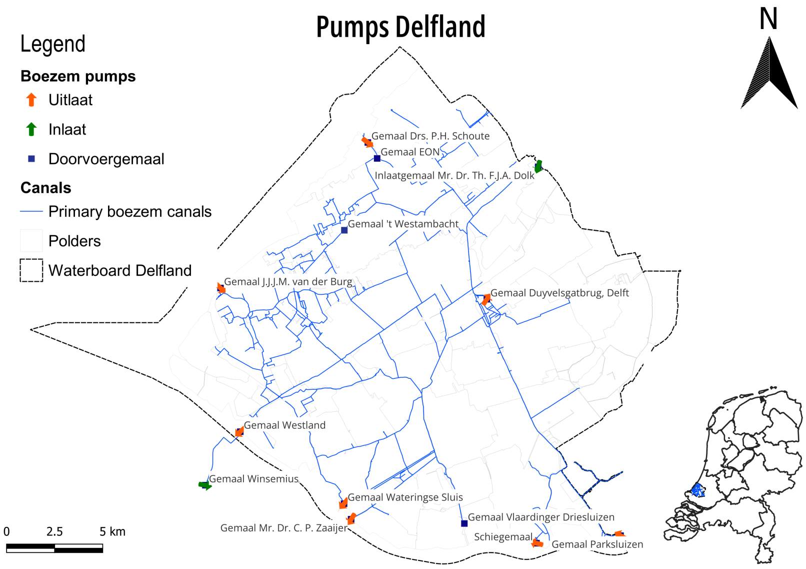


Figure 2.4: Hoogheemraadschap Delfland overview 'boezem' pumps Delfland, 2022

2.2.3. Decision Supportive System (DSS)

The water level in the boezem is being monitored and controlled by the Decision Supportive System (DSS), in Dutch refered to as BOS ('Beslissings Ondersteundend Systeem'). This system operates the pumps in case of a 'normal situation'. The DSS runs different tasks, such as data importation and pre-processing. The DSS is directed by a Flood forecasting and Early Warning System, FEWS. Different hydrological models use this data to make a prediction of the amount of water that needs to be pumped out of the system and which pumps need to operate in order to achieve this. The Delft-FEWS software has a flexible and modular structure, hence FEWS is used for various applications, such as day-to-day operational management and real-time control ("About Delft-FEWS - Delft-FEWS - oss.deltares.nl", n.d.). The RTC-tools optimization is the 'Real Time Control' toolbox that is used to run the Sobek RR model and the data-preprocessing. The Sobek RR (Rainfall Runoff Open Water) model, is a physical model that makes an estimation of the amount of water by schematizing the catchment areas as small 'buckets'. A schematical overview of the inputs and models used in the DSS is presented in Figure 2.5. The Sobek RR model, RTC-tools optimization and the FEWS system are all developed by Deltares. The set-up of Sobek RR, RTC-tools and FEWS is done by HKV and Witteveen+Bos.

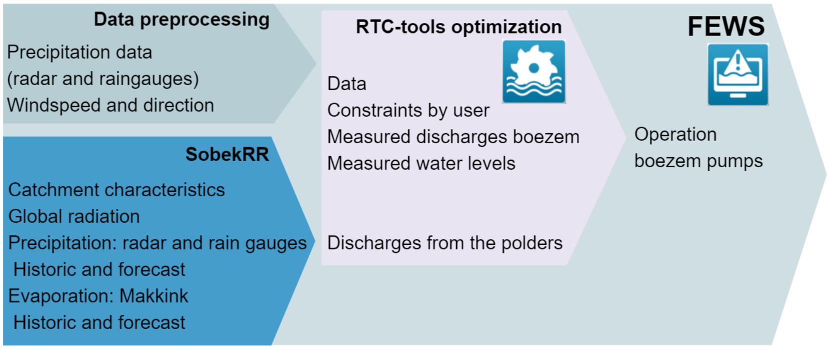


Figure 2.5: Inputs and models FEWS (current Decision Supportive System)

An important historical input is the radar data of the precipitation, this is obtained from HydroNET, which is a product of Hydrologic. The radar data is obtained via the KNMI stations and has a temporal resolution of 5 minutes. For the control the real time observations are used, since the final product of the radar data is only available much later. Next to the radar data, the observations obtained from the rain gauges are used in the FEWS system to make an interpolation for each polder area based on the closest measurement station (Thiessen interpolation). The historical daily evaporation data from the KNMI of the nearest measurement station (Hoek van Holland or Rotterdam) is used for every polder. The precipitation data is resampled to hourly data and used for the calculation of the historical state of the Sobek RR model and the RTC-tools optimization.

For the radar and evaporation predictions the weather model of the KNMI 'HARMONIE' (HIRLAM ALADIN Research on Mesoscale Operational NWP in Euromed, KNMI, 2022) is used. These predictions are updated every hour and provide insights in the precipitation in the next 6 hours per polder area. The predictions of the evaporation and the precipitation data have an hourly resolution and give information up to 6 hours ahead.

The water level in the boezem is being controlled and optimized by an RTC-tools model, based on predictions of the discharges from the polders in the Sobek RR model. In this model the inputs are translated to pumped discharges from the polder pumps from and to the boezem. The global radiation and the Makkink evaporation are used to make an estimation of the amount of water that will be used for evaporation and for the transpiration of plants (in a greenhouse or in an agricultural field). The RTC-tools model uses predictions and measured values of the discharge from the polders to the boezem and determines which boezem pumps should be turned on. The prediction of the RTC-tools model is optimized based on 36 hours ahead and runs every 10 minutes.

In most of the cases, the DSS operates using the optimized configuration of RTC-tools. It is possible to put some 'constraints' on the use of some boezem pumps. For example if there is a high risk of salt intrusion, the system is told to avoid using a specific pump which pumps the water into the system with a higher salt-concentration. The 138 pumping stations in Delfland are simplified into 25 pumps to make it possible to calculate the discharges 'faster'. The current system of the simplified DSS is presented in the figure below (Figure 2.6).

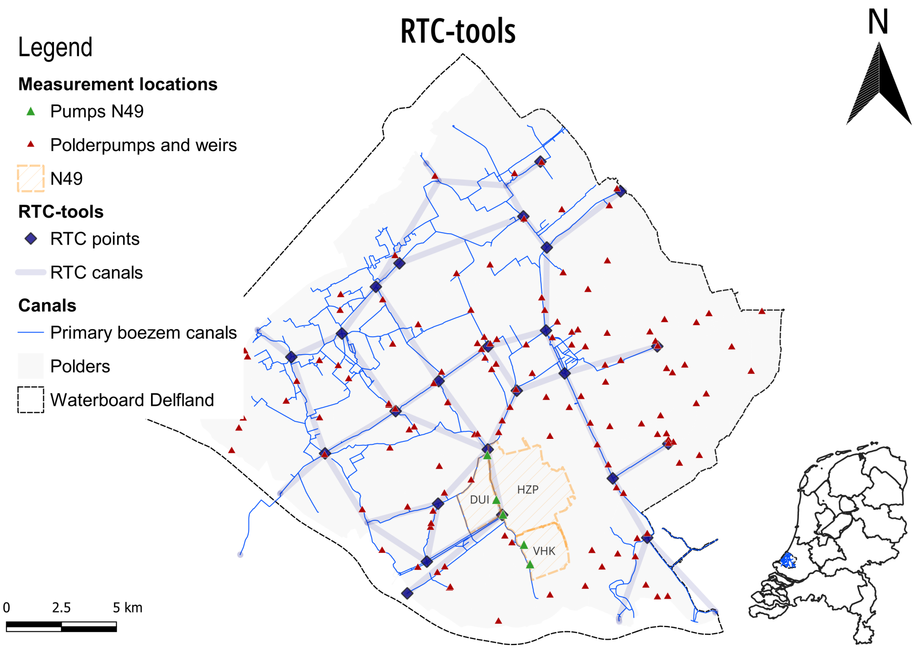


Figure 2.6: Map of Delfland showing the primary boezem canals, boezem canals and the polder canals. The simplified structure of the RTC-tools model is presented in lightblue and the red triangles represent the polder pumps. (“ArcGIS Hub”, 2021 PDOK, 2021 “Bestand bodemgebruik”, 2017 Delfland, 2022
)

2.3. Hydrological modelling

The three main types of models for simulating rainfall-runoff responses are conceptual models, physically based models, and fully data-driven models (Kratzert et al., 2018). Conceptual and physically-based models are typically based on the use of empirical and analytical formulas grounded in the understanding of physical processes (Lü et al., 2012). These models are often referred to as process-driven models.

In contrast, data-driven models operate by capturing the complex relationships between meteorological data and runoff, without the need of explicit knowledge of the underlying physical mechanisms of the hydrological system (Kan et al., n.d.). Understanding of all the complex physical mechanisms and relations is not necessary to make a prediction. The advantage of data-driven models is that the computational time of this type of models is shorter, because it only uses data as input and has no ‘physical constraints’.

There are a lot of data-driven models available. For machine learning (ML) models the algorithms are designed to learn and make predictions or decisions based on features extracted from the data. Examples of ML models are linear regression, random forests or support vector machines (SVM). Feature engineering is an important step for these type of models. A deep learning (DL) model is a specialized form of machine learning which relies on neural networks with multiple layers to learn directly from data. Each layer processes the information and passes it to the next layer for further abstraction. A deep learning (DL) model automatically learns and extracts features from the data, eliminating the need for manual feature engineering. This makes DL models more computationally intensive and less interpretable than ’normal’ ML models. More information on the ML and DL in general can be found in Appendix A.2.

2.4. Random forest model

Random forest (RF) regression is a popular machine learning method for making predictions. The advantage of this method is that the model is able to work with large datasets and can handle also incomplete datasets (Breiman, 2001). It is fast in training and evaluating and can be used to solve both classification and regression-based problems. Another advantage is that a RF model is robust to outliers and can find straightforward and more complicated linear and nonlinear relationships in the data (M. Li et al., 2020) (Breiman, 2001). The RF regression is a tree based algorithm, which is in theory interpretable. The big advantage of RF models is that it is possible to get insights in the features that are the most important for the prediction of the objective variable. A random forest uses a technique called ‘ensemble learning’, which means a prediction is made based on a combination of the predictions of multiple machine learning algorithms (M. Li et al., 2020). In this method the output is the average of all the different outcomes, which avoids that the prediction is influenced by individual errors.

2.5. LSTM model

2.5.1. Neural networks

Neural networks are inspired by the structure and functioning of biological neurons in the human brain. They consist of layers of interconnected nodes (neurons) that process information. Deep learning models are neural networks that have two or more hidden layers. A deep learning model is a more advanced type of machine learning that automatically learns complex patterns from data. The layers are typically capable of learning through training, and often employ non-linear functions to map inputs to outputs. This is done by using back-propagation algorithms to indicate how internal parameters are changed to compute a representation based on the previous layer (Lecun et al., 2015).

2.5.2. RNN and LSTM

Recurrent neural networks (RNNs) is one of the types of Neural Networks that are able to process sequential data, such as text and voice. The disadvantage of using a RNN is that the training of the network is complex. The most popular type of RNNs is a Long Term Short Memory (LSTM) network, suggested more than 25 years ago by Hochreiter and Schmidhuber (1997). It was designed to overcome the weakness of traditional RNN to learn long-term dependencies. It can find long-term dependencies between the input and output data. In the subsequent years the LSTM networks have been applied numerous times, for example in the field of speech recognition, traffic forecasts and for estimation of the fluctuation of stock prices (Song et al., 2020). The LSTM has been proven to perform similar or even better than physical models for rainfall-runoff modelling (Kratzert, Klotz, Shalev, et al., 2019, Frame et al., 2022, Shi et al., n.d. and Hao and Bai, 2023).

2.5.3. LSTM structure

A LSTM network consists of three parts, the forget gate, the input gate and the output gate. An LSTM has a hidden state of a previous time stamp and also the hidden state of the current time stamp, which presents the short-term memory (bottom line). Next to that, there is a previous and current cell state present, which is known as the long-term memory (top line).

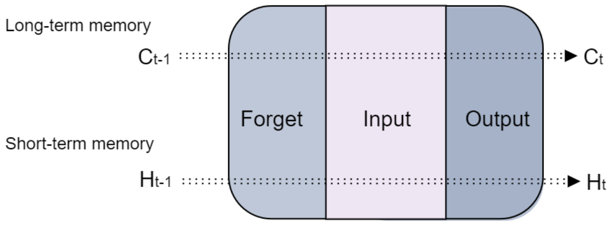
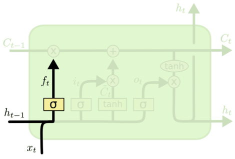


Figure 2.7: LSTM cell

In the first part, the forget gate, a decision is made whether the information coming from the previous time stamp should be forgotten or should be remembered. In this problem the previous time stamp is the observed value of 15 minutes ago. The equation for the forget gate:



$$f_t = \sigma(x_t * U_f + H_{t-1} * W_f)$$

x_t = input to the current time stamp
U_f = weight associated with the input
H_{t-1} = The hidden state of the previous time stamp
W_f = It is the weight matrix associated with the hidden state

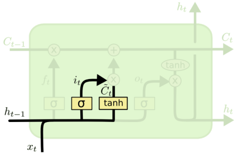
The next step is to apply a sigmoid function (σ). This will result in an outcome between 0 and 1. This number is multiplied with the cell state of the time step before C_{t-1} (long term memory).
–> In the case that f_t=0, everyting will be forgotten.

$$f_t * C_{t-1} = 0$$

–> If f_t=1, the information of the previous step will be kept.

$$f_t * C_{t-1} = C_t$$

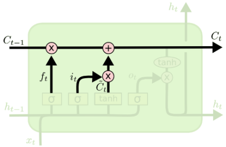
In the second cell, the input gate, the input is processed to learn from it and to use it as new information. The input gate consists of two layers, a sigmoid layer and a Tanh layer in which the new values are created. The input (h_{t-1} and x_t) will be converted by the activation function (sigmoid) to a value between 0 and 1. Then the new information, which depends on the information of the hidden state and input x. This new information will be given a value between -1 and 1 with tanh, if the value is negative, the new information is removed from the cell state and if the value is positive, it will be added to the cell state of the current step.



$$i_t = \sigma(x_t * U_i + H_{t-1} * W_i)$$
$$C_t = \text{tanh}\,(x_t * U_c + H_{t-1} * W_c)$$

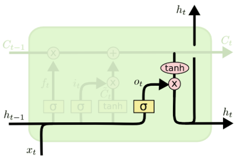
x_t = input to the current time stamp
U_i = weight associated with the input
H_{t-1} = The hidden state of the previous time stamp
W_i = It is the weight matrix associated with the hidden state

Next, the new cell state is calculated.



$$C_t = f_t * C_{t-1} + i_t * {}_t$$

The last cell is the output gate, in this cell the updated information is passed to the next step. Next, the current hidden state is calculated by multiplying this value by the tanh of the cell state.



$$o_t = \sigma(x_t * U_o + H_{t-1} * W_o)$$

$$H_t = o_t * \text{tanh}(C_t)$$

x_t = input to the current time stamp
U_i = weight associated with the input
H_{t-1} = The hidden state of the previous time stamp
W_i = It is the weight matrix associated with the hidden state

3 Material and methods

The methodology used to address the research questions is split into three parts. First, the data preparation and preprocessing are discussed. Then the set-up of the models is discussed, covering aspects such as hyperparameter tuning, data splitting, and sequence generation for the deep learning model. Next to that, the benchmark models selected for comparison are discussed and the methods for evaluation. An overview of the different steps that have been taken is presented in Figure 3.4. In the third part, the approach for researching the attitude towards deep learning in the organization is discussed. This includes outlining the specific questions posed during interviews with representatives of the organization.

3.1. Study area

In this research the focus will primarily be on the examination of a specific polder in Delfland, known as the 'Duifpolder' (DUI), see Figure 3.1. Its primary function is agricultural, consisting of extensive grass and corn fields, separated by polder canals.

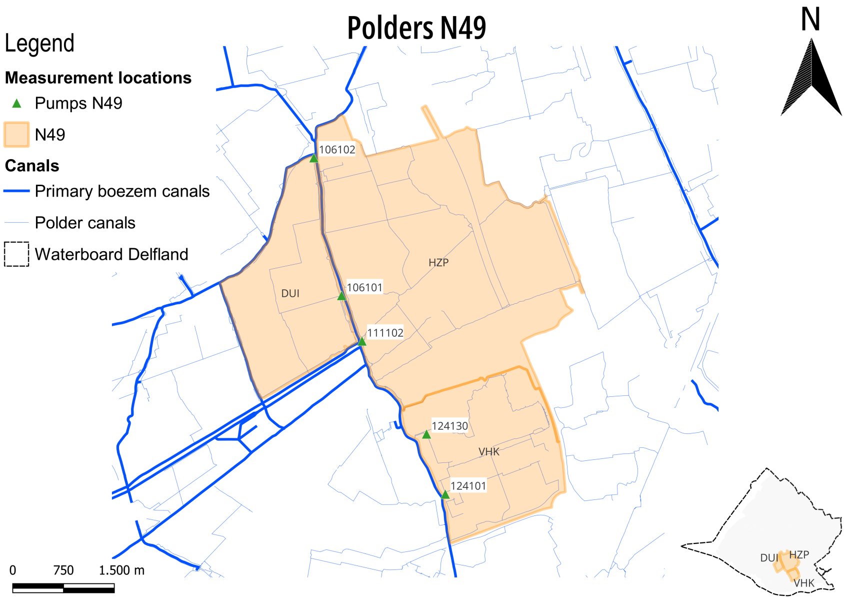


Figure 3.1: Node 49 RTC-tools (study area)

The selection of the Duifpolder is grounded in the availability of comprehensive and complete time series data starting in 2014 for both water levels and discharge. The system did not change in this period, which makes it a useful case study for a machine learning model. Additionally, the two pumps

located in the Duifpolder are part of node 49 (N49), encompassing only four polder pumps. Consequently, this necessitates the development of four distinct models, the outcomes of which will be aggregated for comparison with the Sobek RR model. The following abbreviations for the pumping stations will be used:
HZP Holierhoekse- en Zouteveensepolder pumping station
DUI1 Duifpolder pumping station 1
DUI2 Duifpolder pumping station 2
VHK Vlaardingen-Holierhoek pumping station

3.2. Data

Delfland will provide the data for this research. An overview of the available data is presented in Appendix A.3. To use the data for any model, several steps must be taken. The steps are presented below:

1. Data preparation
2. Data preprocessing
3. Features

3.2.1. Data preparation

The data preparation are the steps which are taken to prepare the data. The first step is the construction of the dataframes of all the polders by combining the time series of Delfland and the KNMI and making sure all the variables are in correct units.

Constructing dataframe

- The observed discharges, water levels and precipitation data of the waterboard of Delfland is collected and from the KNMI the daily radiation, evapotranspiration and temperature data is collected.
- The KNMI station of Hoek van Holland is chosen as reference for all the polders. This is done because the differences between the stations in and around the area are small, as discussed in Appendix A.3.3.
- The data of the discharges, water levels and precipitation need to be connected to the correct polder pump.
    - The coupling of the radar data of Delfland to the pumping station is done by using the area of the polder in which the pump is operating.
    - In some polders, there are two or three pumping stations operating. For example, in the case of the ‘Duifpolder’, in which two pumps are operating. This means that the measured discharges and water levels are different for the two locations.
    - The polder pump is coupled to the nearest rain gauge station from the polder pump, because this will be representative for the total amount of precipitation in the area. In total there are 10 rain gauges in the area of Delfland (see Appendix A.3)
- The data needs to be in a similar format because the measurements of different variables are measured with different time steps. For example, some measurements are made every 5 minutes, and others have a temporal resolution of 15 minutes. The data collected from the KNMI (KNMI, 2022) has a daily temporal frequency, this means it needs to be resampled to measurements every 15 minutes using interpolation. The type of interpolation that is done is linear interpolation, meaning that it is assumed that the value changes with a constant rate over the time. This results in a dataframe of 96 values instead of one value every day.
- The outliers and missing values are removed from the data.

The dataframe consists of the following features:

- Water level [m NAP]
- Discharge [m^3/15 min]
- Water level $\frac{dy}{dx}$ [m/h]

- Rainfall intensity from radar [mm/ 15 min]
- Rainfall intensity from raingauges [mm/ 15 min]
- Temperature (TG) [°C]
- Global radiation (Q) [J/cm^2]
- Mean relative atmospheric humidity (UG) [%]
- Potential evaporation (Makkink) (EV24) [mm]

In some cases there are two or three pumps operating, which means that the varying locations need to be taken into account. Next to that, the different time series that are added together are collected from different sources, meaning that they do not always have the same start and end point.

3.2.2. Data preprocessing

In the previous step the dataframe is prepared. The next step is to transform this dataframe with data preprocessing to a format that can be used for a machine learning algorithm, this means additional features are added and the data is scaled and split into training, validation and test data.

Inlets

If the pump is inactive, the water level in the polder canals increases most of the time. The increase in water level is not caused exclusively by precipitation entering the system. Other triggers are the inlet of water from other polders or from the boezem canals. These inlets take place at a lot of locations and are not measured by Delfland. The inlets are estimated by setting up a water balance by taking into account all the incoming and outgoing water in the polder, as schematized in Figure 3.2. 3.2.

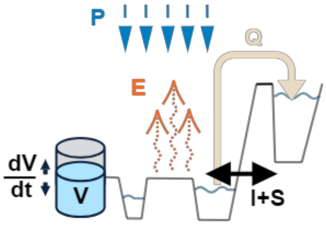


Figure 3.2: Water balance in the polder

The water balance in a polder is calculated using the formula below:

$$\frac{dV}{dt} = I + S + P - E - Q_{\text{pumped}} \tag{3.1}$$

Parameters
$\frac{\text{dV}}{\text{dt}}$ = Change in storage
I = Inlet (unknown amount of water that enters the system) [mm]
S = Seepage of water from other polders [mm]
P = Precipitation [mm]
E = Evapotranspiration (Makkink, KNMI Hoek van Holland) [mm]
Q_{pumped} = The total amount of water that is pumped out of the system [m^3/min] converted to [mm] using the surface of the polder and 15 minutes in each time step.

Features

An overview of all the time dependent features is presented in Table 3.1. The data is obtained from the KNMI and Delfland. In Table 3.1 the time interval, the unit and the source is presented. Some of the features are not specific per polder and the same for all dataframes, such as the value for evapotranspiration. In some cases, this 'simplification' is because there is no data with a better spatial resolution available.

The additional features are the features that are extracted from these data. These features could improve the predictive power of the data. For example, the additional features ‘time of day’ and ‘day of the year’ can give some insight in which season it is and if it is night- or daytime. This is done by converting the number of the day (1-365) to a value between -1 and 1 using a cosine and sine. In this way the last days of the year are close to 1 and are recognized as being the days before the first days of the year (which are also close to 1 due to this transformation).

Another feature that is added is the number of time steps (of 15 minutes) without pumping. This increases every time step when the pump is not switched on. Additionally, the time steps with no rain is added; this contains information about the period without rain, considering the radar data.

The prediction of precipitation for the next three hours is included as a feature. This will provide information about the expected amount of water entering the system; this information is particularly valuable when estimating cumulative discharge over an extended future time frame. Given the absence of historical predictions, observed precipitation data serves as a ’prediction’. For the prediction, the rain gauge data has been used. It is important to acknowledge that this represents the best possible estimation of future precipitation, and in reality this estimate will not be as accurate.

Table 3.1: Features

Category	Feature	Unit	Time interval	Source	Polder specific
Climate	Precipitation deficit	mm	Daily	KNMI	No
	Temperature	°C	Daily	KNMI	No
	Potential evaporation (Makkink)	mm	Daily	KNMI	No
	Global radiation	J/cm^2	Daily	KNMI	No
	Hours of sunshine	hours	Daily	KNMI	No
Precipitation	Radar	mm	5 minutes	Hoogheemraadschap Delfland (HydroNET)	Yes
	Rain gauges	mm	15 minutes	Hoogheemraadschap Delfland	Yes
	Prediction* (next 3 hours)	mm	15 minutes	Hoogheemraadschap Delfland	Yes
Other	Water levels	m N.A.P.	5 minutes	Hoogheemraadschap Delfland	Yes
	Discharge	m^3/min	5 minutes	Hoogheemraadschap Delfland	Yes
Additional features	Day of the year	[-]	Daily	-	No
	Hour of the day	[-]	Hourly	-	No
	Season (0 for winter, 1 for summer)	[-]	Daily	-	No
	# Time no pumping	days	15 minutes	Hoogheemraadschap Delfland	Yes
	# Time no rain	days	15 minutes	Hoogheemraadschap Delfland	Yes
	Inlet (water balance)	mm	15 minutes	Hoogheemraadschap Delfland	Yes

*Observed values of the rain gauge data are shifted to employ as ‘historical predictions’

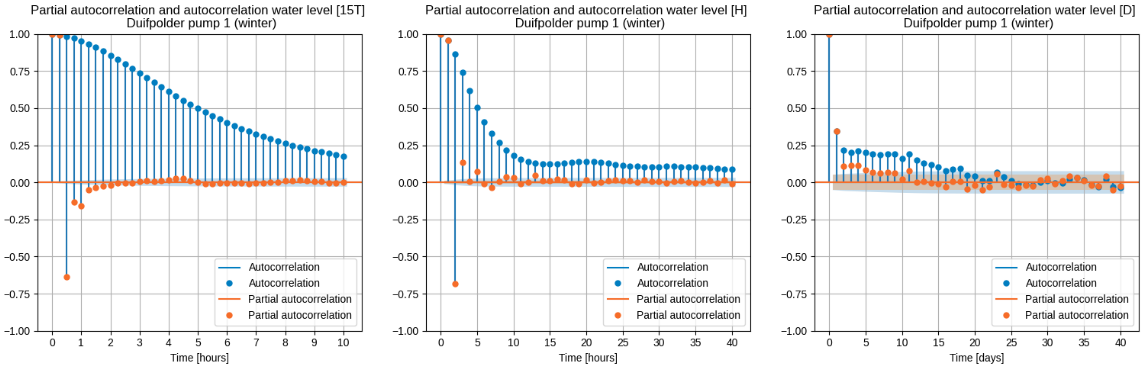
Data analysis

Concentration time The hourly discharge and the different concentration times of the hourly sum of the rain gauge and radar data of the three polders will be analyzed in a random forest regression model. With this model it is possible to obtain insights in the time it takes before the rain ends up in the canals and is consequently pumped out of the system. This analysis is done by using the observed time series in the winter period. This is done because in the winter period the amount of water that is let into the polder is almost zero, which will minimize the disturbance of the correlation between these two variables. With this short analysis, it is possible to gain insight in the important values for the concentration time of the precipitation data.

Auto-correlation and partial auto-correlation water level An auto-correlation plot describes how measurements are correlated looking at different time lags at a certain time step in the time series. The graph ranges from -1 to 1, and if the value is close to 1, it means that there is a high correlation between the two observations. If the value of the auto-correlation is close to 0, it means there is (almost) no correlation between the observations. In the figures for the auto-correlation, also the confidence intervals are presented (blue shaded area). These are set to 95% and if the value of the auto-correlation plots outside of these intervals, it means there is a significant auto-correlation (Valenzuela et al., n.d.). If the auto-correlation remains very high for a long time, this means there is a long-term storage in the system. With an auto-correlation plot it is possible to find patterns in data.

The partial auto-correlation is the correlation between an observation at a given time stamp and the lagged values. This is different from the auto-correlation, since the influence of the intermediate lags is removed. The result is that the focus is only on the direct relationship of the two observations (the data-point and the lagged value).

If the objective is to predict the water levels, it's worth noting that water levels typically exhibit minimal variations over short time intervals. Specifically, when forecasting the water level for a short time horizon (less than one hour), the current water level serves as a highly accurate estimate. This observation is also evident in auto-correlation plots. The high auto-correlation values arise from the consistent behavior of water levels, which typically stay within a defined range without abrupt and extreme fluctuations. Various plots have been generated to visualize the auto-correlation and partial auto-correlation patterns across different time intervals, such as 15-minute, daily, and monthly resampled dataframes.



(a) Resampled dataframe of 15 minutes

(b) Resampled hourly dataframe

(c) Resampled daily dataframe

Figure 3.3: Water level autocorrelation and partial autocorrelation function plots for 15 minutes, hourly and daily resampled time series in the Duifpolder pumping station 1.

Split data

To make a good estimation of the performance of a model, the data on which the model is evaluated, should be 'new' and not seen before, as in a real situation when a model is used. That why the model is split into training, validation, and test data. The training set is used to train the model, the validation set is used to tune hyperparameters, and the test set is used to evaluate the model's performance. Overfitting is reduced since one can evaluate how the model is performing based on the training and test dataset. If the model is doing very well on the training dataset and not on the test dataset, this is a sign of overfitting. There is data available from 10-10-2014 until 10-10-2022, meaning eight complete years of data.

Scaling

Another pre-processing step is the scaling of the features, which means that all features are on a similar scale. This helps prevent features with larger magnitudes from dominating the learning process in the model training and evaluation process. The features are scaled between 0 and 1 with a MinMaxScaler from the package ("1.13. Feature selection — scikit-learn 1.2.2 documentation", n.d.). This is done by subtracting the minimum value of the dataframe (df) and dividing by the difference between the highest and the lowest value of the dataframe (Lecun et al., 2015). It is important that this is done for the columns separately.

$$df_{\text{scaled}} = \frac{df - df_{\text{min}}}{df_{\text{max}} - df_{\text{min}}} \tag{3.2}$$

Another important point that needs to be taken into account is that the scaling is done on the training set only. This scaler can be used for the validation and test set. This is because if the whole dataframe is considered, this means that the model has seen the values of the future during training. The scaler of the training dataframe is used for the validation and test dataframe as well. The scaled data is only used for the LSTM model, since it is required for these type of models in order to avoid that some features dominate the predictions because of the larger magnitude.

3.3. Machine learning models

There are two models used to predict the pumped discharges of the pumps. First of all, a random forest model is constructed to predict the sum of the discharge in the coming 2, 8 and 12 hours. The second model is a LSTM model that will be used to predict the sum of the discharge in the next 12 hours. In this section an overview is presented of the different steps that are taken is presented and the output variable of the model is discussed. Later the development of the two models, including the hyperparameters and features, are discussed. Both the models have been set-up and optimized in a similar way. A schematic overview of the steps is shown in Figure 3.4.

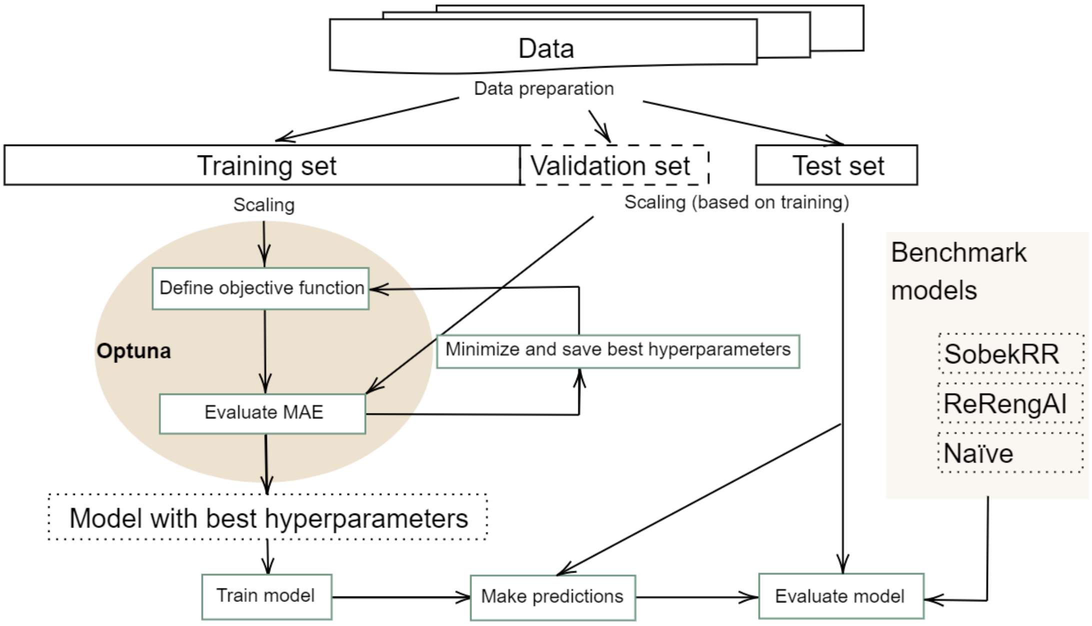


Figure 3.4: Flowchart of the model set-up

3.3.1. Output variable

The objective is to estimate the amount of water from the polders to the boezem canals. It is possible to use the water level as the output variable of the model since the polder pumps operate on the basis of the water level. This means that using the predicted water level the discharge of the pump can be calculated. Another option is to directly use the discharge as the output variable for the model. The pumped discharges exhibit a stepwise pattern that will be more difficult to predict. The primary goal is to forecast the volume of water discharged from each pump in the following hours. The emphasis lies on predicting the total volume of pumped water, rather than the specific operational timing of the pumps. Consequently, the model aims to predict the cumulative discharge within the defined timeframe, meaning that the total water volume is prioritized over pump operation timing.

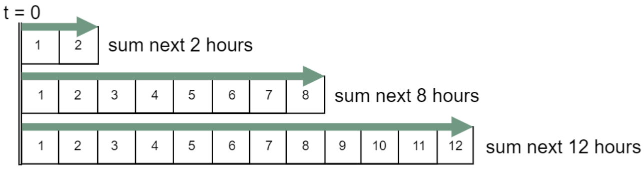


Figure 3.5: Output variable: predicted discharge in the next 2, 8 and 12 hours.

The result of how the gradients of these output variables look for pump 1 of the Duifpolder are shown in Figure 3.6.

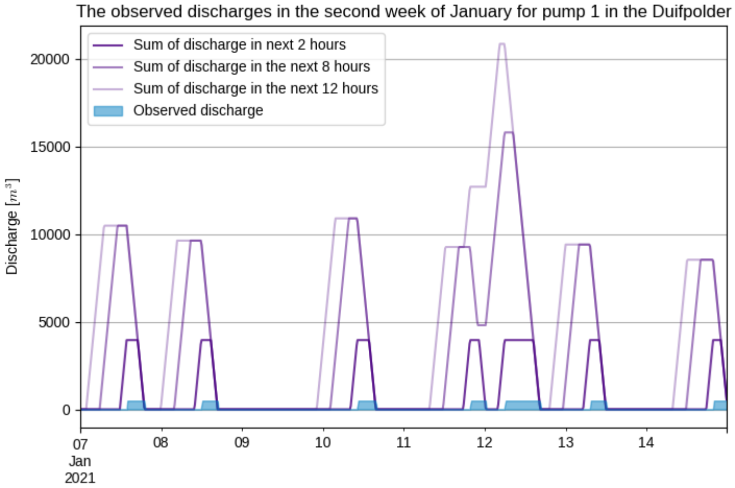


Figure 3.6: Observed discharge and the observed discharge in the next 2, 8 and 12 hours in the first two weeks of January for pump 1 in the Duifpolder.

3.3.2. Model development

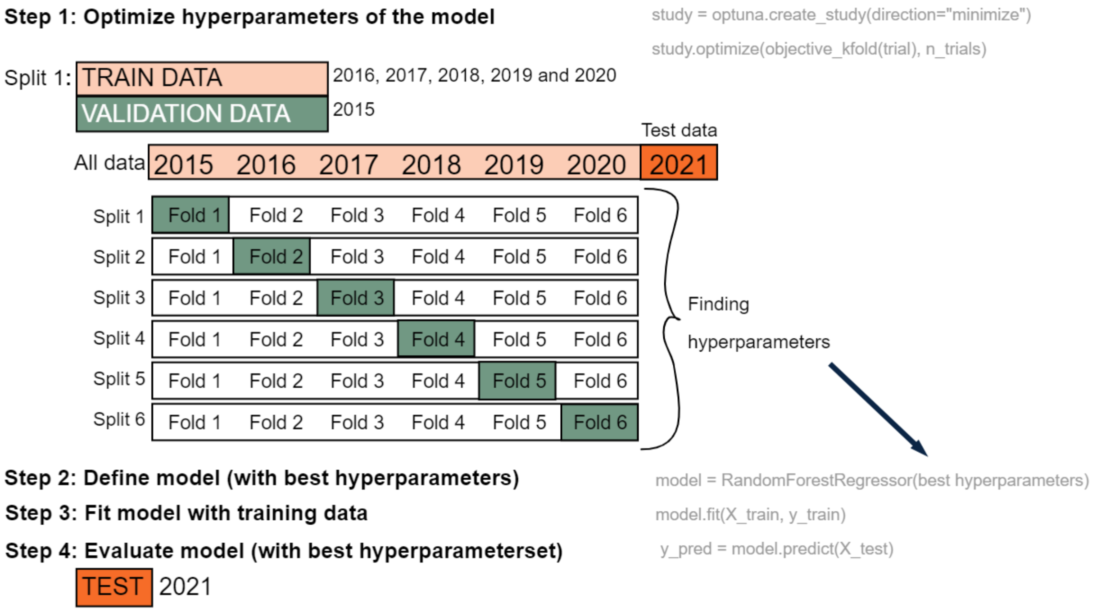
Model 1: random forest regression

With the random forest regression model the sum of the discharge in the next hours is predicted. This value is calculated from the observations and shifted to the 'current' time step. A schematic presentation of what this variable looks like is presented in Figure 3.5.

The steps that have been taken to construct and evaluate the random forest model are discussed in Appendix A.4.1.1. For every polder pump in Node 49 and for the three different time horizons, a model has been set-up. This means in total 12 models are constructed and optimized. Depending on the number of hours that is predicted and which pump is considered, the best hyperparameter set and the feature importances are different. In the results the differences in features and hyperparameters will be researched and discussed.

Feature Engineering In order to make an efficient model, it is beneficial to make a selection of features that have the biggest impact on the model results. The third subquestion is about finding these features and about the different order of importance for the various models. Next to the features that are observed at the time step the discharge need to be predicted, more features can be added, for example the lagged values or the rolling sum of these, to give the model some information about the previous conditions. The most important features are selected and used as input for the LSTM model.

Hyperparameter tuning The model will perform better if the correct hyperparameters are chosen. This is done using k-fold cross validation. In this case the data is also split in the same training and test data, but during the optimization, different 'splits' are considered. K-fold cross-validation is a method that can be used to evaluate the performance of a machine learning model based on choosing another subset of the data for training and testing for every fold. In Figure 3.7 the way the data is split to perform the k-fold cross validation is shown. The training and validation data is split again in six sets of one year. Each time one year is used for the evaluation of the trained model and the other 5 years are used for the training of the model. This means the hyperparameters are evaluated six times based on every split and the outcome is the average Mean Absolute Error of all the splits. In this way overfitting on the training data is reduced and the model is more robust. Normally for timeseries, this way of splitting is done in a different way in which the future data cannot be part of the training set. In this situation the time dependency is low, meaning that an observed value in the next year, does not provide information for the current year. Because there is high frequency data available and the system is currently brought back to the same 'state', the future data in this specific case can be used as training for previous years.


Step 1: Optimize hyperparameters of the model
Split 1:
TRAIN DATA
2016, 2017, 2018, 2019 and 2020
VALIDATION DATA
2015
Test data
All data
2015
2016
2017
2018
2019
2020
2021
Finding
hyperparameters
Step 2: Define model (with best hyperparameters)
Step 3: Fit model with training data
Step 4: Evaluate model (with best hyperparameterset)
TEST
2021

Split 1	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6
Split 2	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6
Split 3	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6
Split 4	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6
Split 5	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6
Split 6	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6

```
study = optuna.create_study(direction="minimize")
study.optimize(objective_kfold(trial), n_trials)
```

```
model = RandomForestRegressor(best hyperparameters)
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
```

Figure 3.7: Steps k-fold cross-validation hyperparameter tuning

The hyperparameters (Probst et al., 2019) that will be tuned are the number of estimators, the maximum depth of the trees, the minimum samples split and the minimum number of samples of a leaf node. A schematic overview of these hyperparameters can be seen in Figure 3.8 and the complete list of hyperparameters is presented in appendix A.4.1.2.

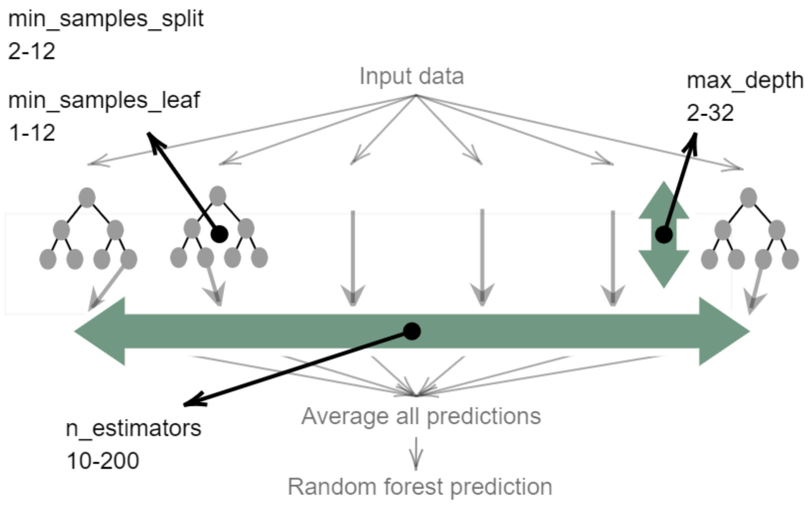


Figure 3.8: Schematic overview of the optimized hyperparameters of the random forest model.

Model 2: LSTM
The Long Short-Term Memory neural network is set up using the Google Tensorflow library (Google, 2023). The inputs, construction of the model and outputs are slightly different than for the random forest regression. It has been chosen to build the LSTM model only for predicting the sum in the next 12 hours because of the long computational time of these type of models. The steps that have been taken to construct a LSTM model are presented in Appendix A.4.2.1.

Features An LSTM model is a complex model consisting of different layers of neurons, as discussed earlier in 2. The training process will take a very long time if all inputs are considered. Based on the feature analysis done with the random forest model, a selection of features is made that will be used as inputs. The input window that is considered is 48 hours, meaning 192 time steps, this number is chosen based on the concentration time of the polders. In about 48 hours 90% of the precipitation has been discharged to the canals (Section 4.3.2).

Sequences A LSTM network uses a sequence of inputs of one or multiple variables and then predicts the future values based on this input. Therefore it is needed to build sequences, which is done with a function that creates these sequences for the train, validation and test data. It is important that the first layer of the model has the same shape as the input. Next to that, the last layer has to be the same shape as the values that are the output of the model. Therefore the input and output can be visualized as presented in Figure 3.9.

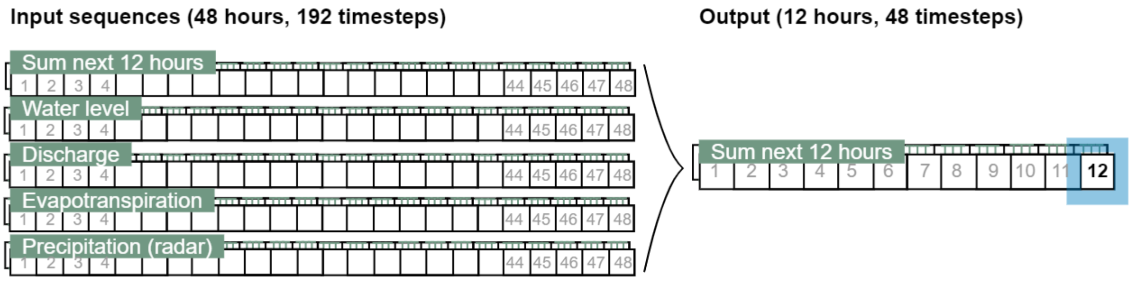


Figure 3.9: Input and output sequences for the LSTM model

Framework The model has been built in Tensorflow - Keras. Keras is part of the TensorFlow platform and provides an easy to use interface for making machine learning models, focused on deep learning models, such as a LSTM network. With the Keras API one can create layers, models and data preproccesing tasks. After the model has been made, the model can be trained with the *tf.keras.Model.fit* method. After the training process, the output of the model can be generated with the *.predict* method. The model can be evaluated with the *.evaluate* method.

Callbacks The process of training a deep learning model can be time consuming. To avoid wasting time on overfitting the data and to stop the optimization process at the right point, it is necessary to implement callbacks. One commonly used method for controlling the training process is EarlyStopping. EarlyStopping is a function of Keras which can stop the training if there are no changes in the loss value after a specified amount of epochs. Another callback that is used, is the 'ModelCheckpoint', this callback is used to store the weights and hyperparameters of the model once the training process is terminated.

Hyperparameters A model architecture for the LSTM model needs to be chosen. The architecture of a machine learning model is determined by the hyperparameters. These hyperparameters need to be optimized since the values are specific for each model. This will be done with an optimization algorithm, called 'Optuna' (Lim, 2022), further discussed in Section 3.3.3. The hyperparameters that are optimized for the LSTM model are the number of nodes, activation function, dropout rate, learning rate and the batch size, further explained in Appendix A.4.2.2. A selection of hyperparameters is made to avoid very long computational times for the optimization.

3.3.3. Optimization model

For a RF model the hyperparameters that are tuned are the `n_estimators`, `max_depth`, `min_samples_leaf` and the `min_samples_ split`. These hyperparameters for a LSTM model are for example the number of neurons, the type of activation function, the drop-out rate, batch size and the learning rate. For every hyperparameter a range is given in between the value has to be. The hyperparameters of a model are important for a good model performance because they define the architecture and controlling the learning process of the model (Nguyen et al., 2020). The best hyperparameters vary for each machine learning task. The search for the optimal set of hyperparameters can be difficult and not very efficient, that is why it can be chosen to do this automatically by a framework. The advantages of an automatic search is that it reduces the human effort, it increases the model performance and it makes it more reliable to compare model results if the models are not manually tuned by humans (Nguyen et al., 2020).

There are many optimization frameworks available and in this study the Optuna framework is chosen (Lim, 2022). Optuna can be used for the optimization for both the random forest and the LSTM model. Optuna has an efficient sampling and pruning algorithm, that can be defined by the user itself. It is possible to manually define the search space for the hyperparameters, which makes Optuna easy to implement (Akiba et al., n.d.).

Optuna is very suitable for larger datasets and more complex models. It is an optimizer that needs an objective function, which returns a value that evaluates the performance of the hyperparameters. The hyperparameters can be integers, floats or lists. An example of the objective function that is used for the hyperparameter search for the RF model is presented in Appendix A.4.3. The function returns the MAE of each hyperparameter trail. To find the best set, the output of the objective function is minimized for a defined number of trails.

The algorithms used for optimization in Optuna can be split into two categories, the sampling strategy and the pruning strategy. The sampling strategy focuses on the areas in which the hyperparameters give the best results and selects that specific parameter combination. The other category is the ‘pruning strategy’, which optimizes based on early stopping (Akiba et al., n.d.).

The Tree-Structured Parzen Estimator (TPE) is the default sampler which uses Bayesian Optimization. This makes the search more efficient compared to traditional optimization techniques, like grid search or random search, by picking new points based on previous good results. The pruning strategy implies that some trials with bad results are terminated. This is called early-stopping and avoids the waste of time on hyperparameters that are giving unpromising results. Consequently, the number of samples taken from the hyperparameter search space is guided by probabilistic principles, leading to a reduction in the overall number of evaluations. This enables a focused assessment of the most promising candidates for hyperparameter selection. Pruning will save time and computational resources by avoiding the exploration of hyperparameters that are unlikely to lead to good performance (Nguyen et al., 2020).

Visualization training process

The training process will be visualized in order to get an idea of how the loss is decreasing over the epochs. This shows if the model is ‘learning’ over time and provides insights in whether a model is over- or underfitting. As the model is evaluated based on the MSE, this loss is squared in order to get an idea of the magnitude. Another important step that needs to be done afterwards, is the rescaling of the errors. All the sequences that are the input of the LSTM model are scaled and need to be scaled back to the ‘normal’ ranges. This will be done with the scaler that is stored at the beginning. The training process will be visualized by plotting the number of epochs at the x-axis and presenting the corresponding loss on the y-axis. This will be done for the training dataset as well as for the validation dataset.

3.4. Evaluation

In this section, the benchmark models used for comparison are introduced. The different models will be compared by computing several error metrics, as presented in the second part of this section.

3.4.1. Benchmark models

Inclusion of benchmark models allows us to contextualize the performance of our models against established references. Given that the model predicts discharge sums for the next 2, 8 and 12 hours, we ensure that the benchmark models’ predictions are also converted to these corresponding time frames.

The benchmark models are:

- ReRengAI

- Sobek RR
- Naïve model

The three types of benchmark models that are used to compare the performance of the RF and LSTM model are the machine learning of Delfland (ReRengAI), process-based model of Delfland (Sobek RR) and the naïve model. The first benchmark model is RerengAI Light Gradient Boosting (LGB) model. This is the in-development machine learning model developed by Delfland (see Appendix A.2), which predicts the discharges for each pump separately. The first six hours of the prediction are based on the LGB-model and the next discharge in the next 30 hours is a simpler multiple linear regression model. The other model that is used as benchmark is the current Sobek RR prediction, which is the physical model of Delfland that is based on equations and conditions. The Sobek model calculates every step using meteorological input data and the physical equations on which this model is based. The model is run at once and is not updated with for example water levels or the times the pump is switched on. The predictions are only available for the year 2021. This model is simplified and the results are only available as the sum of the discharges from several polders at once. The Duifpolder is located in node 49, near the 'Holierhoekse- en Zouteveensepolder' and the 'Vlaardingen-Holierhoek'. These three polders are part of a cluster of pumps. These three polders have in total five polder pumps that are schematized in one RTC-node (N49). One of these pumps is a circulation pump (Dutch: 'circulatiegemaal'), which means that this pump is not discharging in the 'boezem' canals. This means, in order to compare the discharge to the Sobek RR model, the discharge of the four polder pumps need to be calculated. The Sobek model has been run for the year 2022. The Sobek model is presented in the number of m^3 that is expected to be pumped from this node. This means it is one value for all the four pumps in the polder. The last benchmark model is the so-called 'naïve' model. This means the last observed value will be used as a prediction. In this case, for example, the sum of the last 12 hours of discharge is the prediction of the discharge in the next 12 hours.

3.4.2. Comparison performance

After training, the performance of the model can be evaluated using the test set. The test dataset is the data from 10-10-2020 until 10-10-2021. This can be done by computing error metrics, like NSE, RMSE, MAE and the R^2. The formulas of these different error metrics are presented below:

(1) Coefficient of determination (R^2), which measures the relationship between predicted and real outputs with values ranging from 0 to 1

$$\mathrm{R}^2 = 1 - \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}(y-\hat{y})^2}{\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}}(y-\bar{y})}} \tag{3.3}$$

(2) Mean absolute error (MAE), which is the absolute difference between the predicted and actual output.

$$\mathrm{MAE} = \frac{1}{n}\sum_{\mathrm{i}=1}^{\mathrm{n}}|y-\hat{y}| \tag{3.4}$$

(3) Mean squared error (MSE), the squared average difference between the predicted and actual value

$$\mathrm{MSE} = \frac{1}{n}\sum_{\mathrm{i}=1}^{\mathrm{n}}(y-\hat{y})^2 \tag{3.5}$$

(4) Root mean square error (RMSE), which calculates the square root of the average of the error squares between estimated and real values

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}(y-\hat{y})^2}{n}} \tag{3.6}$$

(5) Nash-Sutcliffe efficiency (NSE) (Nash & Sutcliffe, 1970), a normalised metric that determines the residual variance (noise) intensity in relation to the computed variability (information) and thus provides

information on the fit of the model to the observed values.

$$\text{NSE} = 1 - \frac{\sum_{\text{i}=1}^{\text{n}}(y - \hat{y})^2}{\sum_{\text{i}=1}^{\text{n}}(y - \bar{y})^2} \tag{3.7}$$

The formula for R^2 includes the sum of real output minus the average of real output and predicted values minus the average of predicted values, all squared, divided by the sum of real output minus the average of real output, squared. The R^2 is useful for the comparison of different models and easy to understand.

The MAE represents a real value of the magnitude of the error, which is useful for interpreting the effect of the error. The MSE emphasizes larger errors because of the squaring. The RMSE makes the MSE interpretable because the output is squared, and the unit will be the same as the actual unit of the output variable.

The NSE is extensively used in hydrological modeling, as it normalizes precision to a more understandable level and is useful for comparing two models. The NSE value can range between -∞ and 1. The closer the NSE is to 1, the better the fit. If the NSE is negative, the mean of the observations is a better fit than the predicted values (Nash & Sutcliffe, 1970).

For this research it has been chosen to optimize the model on the MAE, since the focus is on the the 'daily' situation. This means that the performance of the model should be best in case of the not extreme situations. In many extreme cases, the operators generally take over control of the model, which is why it is important that the model is optimized based on the 'normal' situation. The training of the model is done on the MSE in order to improve the model performance by 'punishing' making large errors. The MSE will improve the model because the larger errors are emphasized. This in combination with the optimization based on the MAE will be favorable for the final estimation of the discharges.

During the training process, the MSE is used. To interpret the magnitude of the error the outcome needs to be The square root of these values needs to be taken in order to be able to interpret these values. In case of the training process for the LSTM, the errors also need to be scaled back. The combination of the two losses will balance the importance of both objectives and avoids overfitting on extreme observations, which might result in big errors in the cumulative value of the discharge.

3.5. Practical implementation of machine learning models at a water authority

In order to find out more about the steps that must be taken to implement a machine learning model for the operational water management in Delfland, several interviews have been conducted. For each person, the interview was slightly different, in some interviews the focus was more on the technical specifications of the model, while in other interviews the emphasis was on the cooperation between parties in the organization. The questions and discussion points are presented in Appendix A.6.1. In the results chapter, the main conclusions that have been drawn from these interviews are presented. The conclusions can be divided into two main categories, one is about the perspective of Delfland on machine learning and the other is about the constraints that the model needs to comply with.

4 Results

In this chapter the results of the research are presented. In the first two subquestions the data is analyzed and the important notices to take into consideration when constructing a model are discussed. In the third sub-question, the most important features for the prediction of the discharge are studied. For example by looking into the concentration time of the polder and the influences of the selected features on the predictions made by the Random Forest model. With this information, the features for the LSTM model will be selected. A model is made for every pump and for the different time horizons. The model architecture of the RF and the LSTM model is optimized by Optuna which searches for the best hyperparameter set based on the given data. In the fourth sub-question, the optimized hyperparameters of the RF model are compared for the different pumps and the training process is visualized. The results of the two different models are compared to the benchmark models in SQ5. In the last sub-question the focus will be on the steps that have to be taken to implement a machine learning model in an organization like the Water Authority of Delfland.

After the answering the research question, some discussion points and recommendation are given for each subquestion separately.

4.1. Important considerations for modelling the discharge

The objective is to estimate the amount of water from the polders to the boezem canals, so the discharge will be the final output of the model. Given that the operation of the pumping stations is determined by the water levels, the first option is to model the water level and then convert these water levels to a modeled discharge. The second option that is considered is to model the discharge directly. First the advantages and disadvantages of the first option will be discussed This leads to the decision to model the discharges directly instead of the water levels, which is the second option. Next to that, the prediction horizon for the discharges from the polders to the boezem is discussed.

4.1.1. Water level as output variable

The water level in the polder canals is measured next to the corresponding pumping station. The water level as a output variable is favorable since all the effects of irrigation, precipitation and the inlets can be seen in this data. In contrast to that, the pumped discharge is a largely binary signal, which cannot always be clarified. As discussed in Section 3.2.2.3 the water level does not change a lot on the short term, which means this information can be ‘learned’ by the model and improve the results. Next to that, the water level mostly stays within two thresholds, which can be added as additional information to the model. Another advantage of modelling the water level is that a LSTM model is more suitable to predict a signal that has a ‘memory’ and influenced by previous conditions in the system.

The relation between the water level and the pumped discharge is that if the water level in the polder reaches a certain stage, the pumping station is switched on, this can be seen in Figure 4.1. In theory, this value can be used to predict exactly when and for how long the pumping station is switched on. The capacity of the pumping station and the difference between the polder level and the water level in the boezem is known, which means that if we know when and how long a pump is switched on, we can calculate the discharge from the polder.

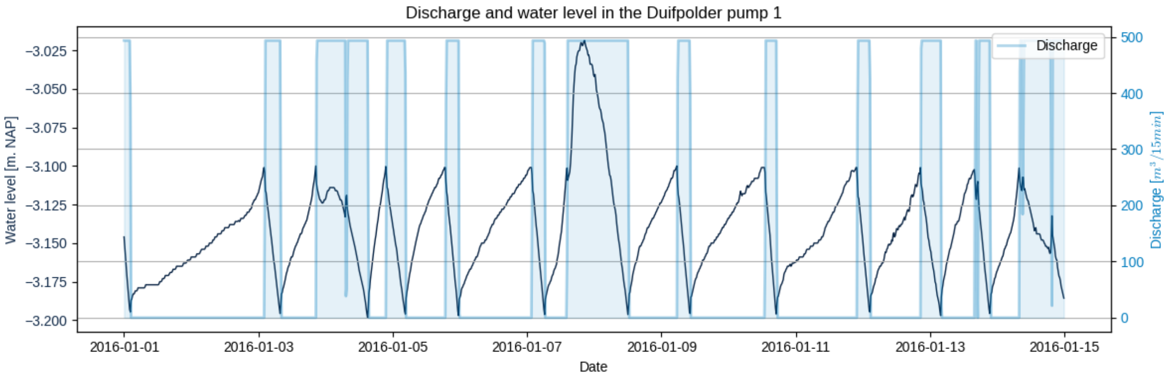


Figure 4.1: Observed discharge and water level in the Duifpolder in the first two weeks of January 2016.

There are several important factors to take into account when predicting the water level in the polder canals in Delfland. First of all, the water level at which the pumping station is switched on or off is different in every polder and sometimes also changes within the polder, if there is more than one pumping station operating in the same polder. Next to that, the threshold water level can change through the year, meaning that there is a summer and winter water level. This water level is determined by the water authority, in the so-called ‘peilbesluit’ (water level ordinance). These water levels are specific for each polder and are based on the different functions that the polder has. For example, in rural areas the water level changes through the year, in summer the water level is higher than in winter times. Nature conservation organizations generally aim for a higher water level, while this can be disadvantageous for the accessibility of the (agricultural) land. This means there can be conflicting interests of the farmers against the nature conservation organizations. Next to that, a sufficiently high water level is also important to avoid damage to the foundations of houses and to minimize soil subsidence (Delfland, 2022). The ‘peilbesluit’ thus changes for each polder and is also sometimes adjusted. For example in the ‘Aalkeet-Buitenpolder’ (ABU), the peilbesluit has been changed in 2018 (see Figure 4.2).

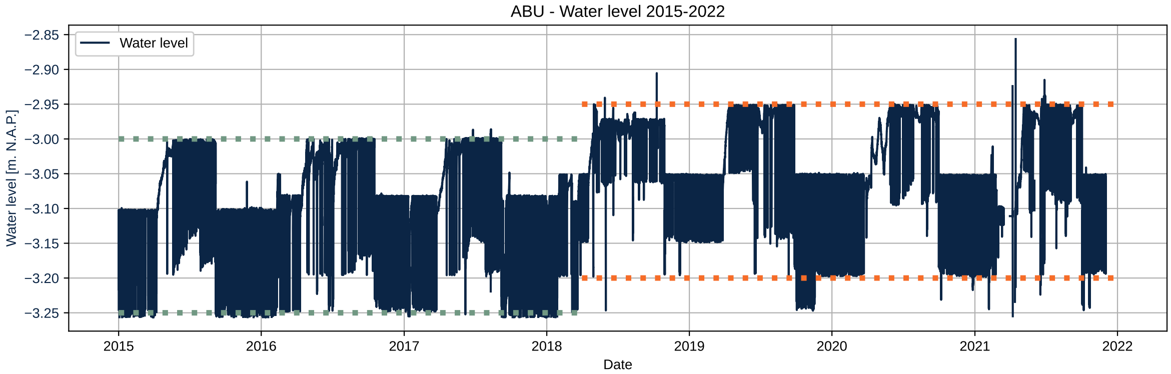


Figure 4.2: Different water level during summer and winter in the Aalkeet-Buitenpolder and a change in water level ordinance in 2018.

Another important consideration for modelling the water level are the human influences, for example the ‘pre-pumping’ (voor bemalen) in case there is a lot of rain coming. Another example of an interference could be irrigation in case of a dry period. In some cases, there are some gaps in the data or there are some outliers, which can complicate the use of this data for a machine learning model. In some cases the water level in the polder canals decreases, without operation of the pump. This happens regularly in for example the Duifpolder. One can see that during daytime the water level decreases and then rises again in the night. This could be explained by farmers irrigating their fields and evapo-

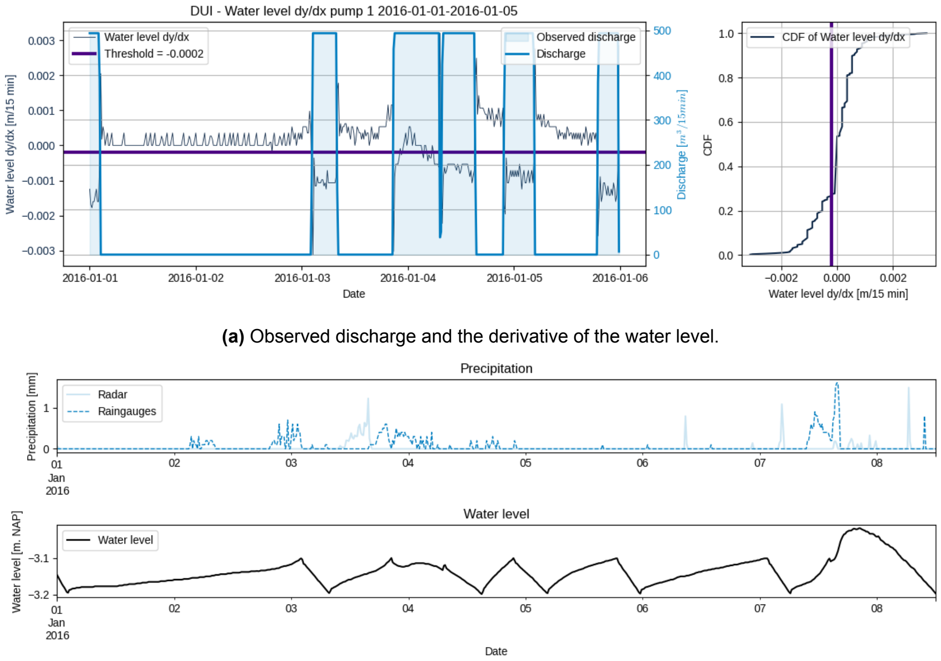
ration. Another option is that the water level decreases because water flows to another polder canal, separated by a weir or gate. The water can also be withdrawed from the polder canal to increase the ground water table at a distance of the canal.
In general, the water level in a canal increases due to precipitation. Though, the water level in the polder can also increase without rain, for example because of not measured inlets in the system from the higher elevated boezem canals, or the seepage from neighboring polders. These inlets will be discussed in more detail in the next sub-question.

Conversion water level to discharge

Once the water level is predicted, it is desired to translate these forecasts into discharge values. The pumping station operates via a 'inslagpeil' and 'uitslagpeil', which means it is automatically turned on and off based on the actual water level. The approach to convert water levels to discharges is studied, which turned out to be not very straightforward. In this section some limitations and possible solutions for the conversion of the water level to the discharge are discussed.

First of all, the pumping station does not always switch off at a certain threshold level (the 'uitslagpeil'). In some cases, for example, if heavy rain is expected, the operators choose to pre-pump some water out of the canals, which will lead to an underestimation of the total amount of discharge. Next to that, these 'inslag peil' and 'uitslag peil' change through summer and winter and also in the different years. This means that using only the water level, the conversion will lead to an error in estimating the discharges, even though the water level is predicted well.

These disadvantages can be 'solved' if the focus is on the derivative of the water level in the polder. If the pumping station is operating, this means the derivative is smaller than zero. To find out whether this conversion would work, an analysis of the time series of the discharge and the derivatives of the observed water levels is done. A figure of these variables plotted next to each other is shown in Figure 4.3, it can be seen that (in general) if the derivative is below a certain (negative) value, it means that the pumping station is operating.



(a) Observed discharge and the derivative of the water level.

(b) Observed precipitation (radar and rain gauges) and water level.

Figure 4.3: Different input features at the location of pumping station 1 in the Duifpolder during the first days of January 2016.

It could happen that the pump is operating, while the derivative of the water level is not below the before mentioned threshold. This can be seen on the day January 4th in Figure 4.3a. This can be explained by a heavy rainfall event in which the pump is operating. The result is the derivative of the water level is above the threshold, while the pump is operating. This means that the conversion of water level gradient into a signal whether the pump is switched on or off can be difficult in some essential cases.

In certain polders, pumping stations operate at varying capacities. Consequently, even if it is known when a pumping station is activated based on water levels, determining whether the pumping station is operating at normal or maximum capacity is not straightforward. This is not a problem for predicting the discharge of the first pumping station in the Duifpolder, since this is always the same. Although, for example, in the Holierhoekse- en Zouteveensepolder and the second pumping station of the Duifpolder the pumping station operates with more than three different capacities (see Appendix A.1.3).

Discharge as target variable

It is not straightforward to calculate the discharges from the water level or the derivative. The difficulties are caused by the pumping stations that operate at different capacities. This means that even if the prediction of whether the pumping station is on or off is known, the crucial information of the capacity at that moment is not obvious. In addition, the process of translating a predicted value into an actual output introduces additional errors into the prediction, as uncertainties exist in both models. That is why it has been chosen to model the discharge directly.

In this case, not the value at a certain point is predicted, but the total pumped discharge in the next 2, 8 or 12 hours. This information can be a useful input to the FEWS model. The observed discharges can be described as a block function. In order to smooth this pattern, the target variable will be the sum of the discharge in the next period. In this way, the focus will be on the amount of water, instead of the exact timing of the blocks. This makes the time series also more suitable for a LSTM model, since there is a short of memory created in the system. This target variable is shown in Figure 4.4. To calculate this sum, the rolling sum of the observed discharge is calculated, and then these values are shifted up in the dataframe, so that the future discharge of a certain period is calculated and can be used as target variable for the models. In the fifth sub-question some experimental results of the LSTM model will be presented for predicting the discharge in the next 12 hours.

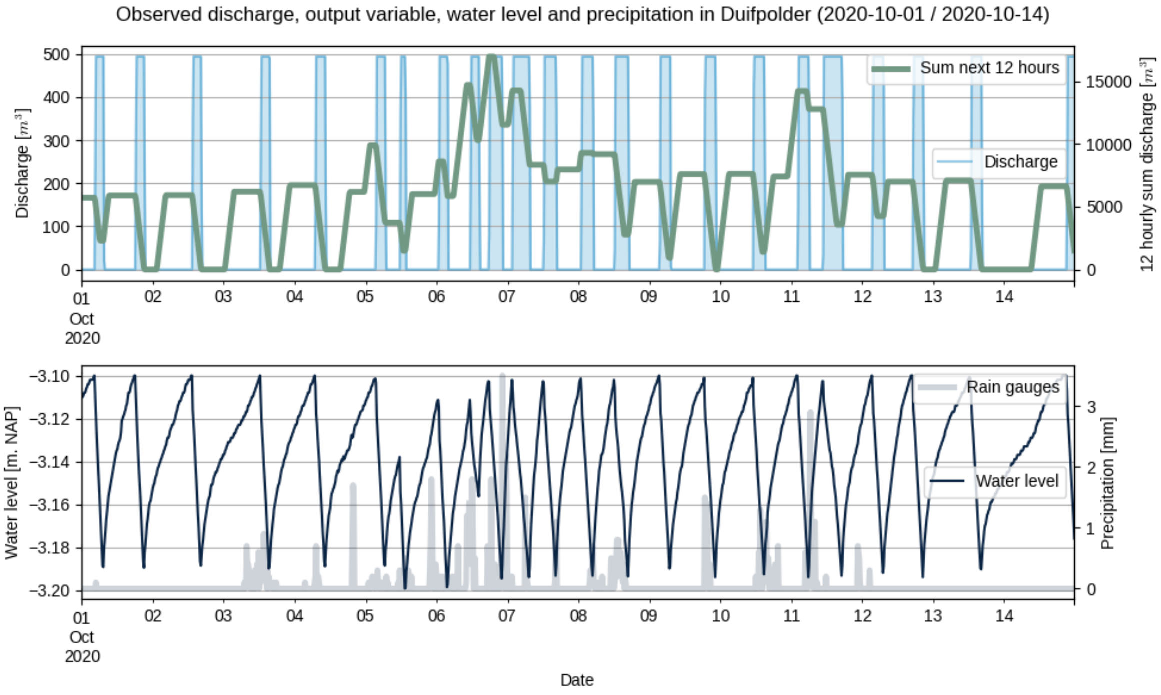


Figure 4.4: Target variable (12 hourly sum of the discharge) in the Duifpolder

4.1.2. Time horizon

In order to know how far ahead the predictions of the discharge should be, some exploration interviews are conducted with representatives of the water authority. In the current model, the discharge from the polders is recalculated every 10 minutes. The Sobek model predicts the amount of water coming from each node separately, which serves as input for the Decision Supportive System (DSS). In this system, it is possible for the operators to see the configuration of the boezem pumps up to 1.5 days ahead. This is updated every 10 minutes and is automatically controlled.

In the current model the predicted discharge of the next 36 hours is used. In the prediction of more than 12 hours, many things can change in the prediction of the precipitation, which is why mainly the prediction of the first 12 hours is relevant for the operators. The prediction of the configuration further in the future is thus only 'theory' and cannot be used since the allowed deviation from the -0.43 m NAP in the system is very small.

It is useful to know the operation of the pumping stations far ahead, but there is a limit that can be done to 'prepare' the system because of the small buffer capacity in the system. In the boezem canals, it is possible to lower the water level about 3-4 cm, since the pumping stations have a sufficient capacity, often the operation starts once it starts raining. In practice, with a good estimation of the discharge in the next 12 hours, this is sufficient for the DSS to respond. That is why it is desired to have a good estimation of the discharges for a time horizon up to 12 hours, so it is decided to predict the discharges up to 2, 8 and 12 hours ahead. The predictions of the discharges calculated with Sobek RR are recalculated every 10 minutes. The new machine learning model can use the past observations of 15 minutes before for the prediction.

4.1.3. Discussion and recommendations

Changes in system

If a data-driven model needs to make predictions for a new situation in which for example the system has changed, the model will not be able to take these changes into account, unless the model is retrained. For example, the water level ordinance can be changed by the water authority. Another property of the polder is the pump capacity. The amount of polder pumps and the capacity per pumping station can be changed if that is desired by the water authority. This can be due to the need for maintenance, capacity increase, or the installation of a new pump. Another change in the system can be caused by pre-pumping. This means that the predicted amount of discharge into the FEWS system is delayed. If this happens regularly in the training data, it might be possible to find this behavior.

A possible suggestion could be to give the model more information on the system, so that this can be taken into account for making the predictions. For example information on the pump capacities, the threshold water levels and the response time of the polder can be added. These information can be used as input, but also to constrain the output space, for example, it is not possible to pump at more than the maximum capacity. Another option would be to constantly check the model inputs. In the case the mean water level increases compared to the average observed water level in the past week, this means the reference water level has changed. If the operator receives a warning about the changed situation that had been induced by himself or a colleague, they can initiate the retraining of the model or changing the reference water level in the polder.

Suggested modelling approaches

Classification Another option would be to make a RF classification model, that only determines whether a pump is switched on or off. Though with this approach the varying pump capacities are still a problem. If a pump operates at for example two different capacities an option to overcome this problem is by extending the classification problem to three classes, off, half capacity or full capacity.

Hybrid Approach In some cases, a hybrid approach that combines the strengths of both data-driven and physically based models may be the most effective solution. This can involve using data-driven models to improve parameter estimation in physically based models.

In practice, the choice between data-driven and physically based models often depends on the specific problem domain, available resources, and the level of understanding required. Many real-world applications benefit from a combination of both approaches to leverage the strengths of each. There are several approaches for combining knowledge and data-driven models, such as ‘Theory-guided data science’ (Karpatne et al., 2017), ‘Informed machine learning’ (Von Rueden et al., 2023), ‘Physics-informed machine learning’ (Karniadakis et al., 2021), ‘Physics-based machine learning’ (Swischuk et al., 2019) and ‘Physics-informed neural networks’ (Raissi et al., 2019).

The physics-informed LSTM neural networks have shown promising results in Rainfall Runoff modelling (Parisouj et al., 2022 and Xiang and Demir, n.d.). In the last mentioned paper, a combination of a LSTM and a Graph Neural Network (GNN) is suggested, this option could be used to obtain a complete overview of the system. This can be favorable for this specific case since some polders interact. It might be an option to create a 2D model for some aggregated polder of the area with this approach.

It was found by Kratzert, Klotz, Herrnegger, et al. (2019) that adding physical constraints to LSTM models could improve the predictions, and it is suggested to look into the possibilities. For example incorporating physical constraints in a custom loss function in which the penalties are bigger once the deviation of the expected behavior based on physics is bigger. It could also be an option to embed the conservation of mass as a constraint to the output space of the model.

Remove time steps while the pump is operating There are several complications when the water level is predicted, since it is not a ‘natural’ response due to the operation of the pump. A possibility is to remove the time steps in which the pump is operating, meaning that the dataset that needs to be predicted would look like in Figure 4.5. In this way only the ‘natural’ response is being predicted.

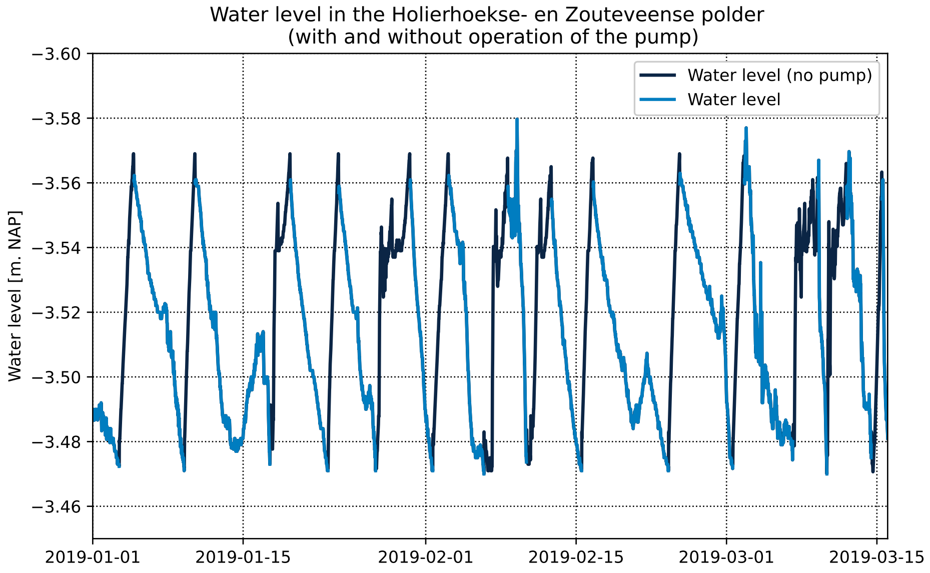


Figure 4.5: Observations of the water level with and without the operation of the pump in the Holierhoekse- and Zouteveensepolder.

Deviation from reference Another option is to take into account the reference level and use it to calculate the difference between the water level and this reference water level. Instead of predicting the absolute water levels the idea is to predict the difference from the reference level. In this way, the change in ’peil besluit’ and summer- and winter water level will be overcome. If this approach will be used, some comments need to be made. First of all, it means the data-preproccessing will be complicated a little bit, since for every polder this winter and summer water level need to be found using the ’peilbesluit’ or by extracting these values from the data. Another drawback of this approach is that if heavy rain is expected and the operator decides to pre-pump some water out of the polder, the water level will be decreased below the set threshold level, which is happening only a few times a year. The predicted values need to be translated to discharges, which influences the final performance of this approach. The advantage of this method is that a model can be made for several polders at the same time, since this additional information that is different for every polder will be taken into account. This way only the response of the system is modelled, and the relation between the forcings such as precipitation and evaporation will be more clear.

Construct model for conversion water level to discharge There are several ways to improve the translation of water levels into discharges. One option is to construct a separate model explicitly for this purpose, this could be added after the model which predicts the water level. Alternatively, another approach involves measuring the water level at two points and utilizing the difference between those points to predict the operating capacity of the pump.

4.2. Estimation of the inlets in the polders from the boezem

Especially during dry summers, a lot of water from the boezem canals is let into the polders. This is done to maintain the water level in the polders and to avoid drying out of the soils. Additionally, the inlets play a role in flushing the polder canals, and thereby improve the water quality. The problem with these inlets is that the quantity is not known. This is because there are several points where these inlets occur. These inlets are impossible to measure for every polder in the area, since these are situated at many points in every polder. The inlets occur sometimes constantly and in some cases occasionally by individuals. In this research question, an estimation of the inlets is made by constructing a water balance. This value is compared to the precipitation deficit in Delfland.

4.2.1. Water balance

In order to get an idea of the quantitative amount of water that can be considered as inlet, the water balance is used. The water balance of each polder is calculated based on the observed discharges (Q) and rainfall (P) obtained from Delfland and the evapotransporation (E) (Makkink, (KNMI, 2022)), see Figure 4.6. The seepage of water from other polders (S) and the amount of water that is let in/out from the polder (I) are unknown.

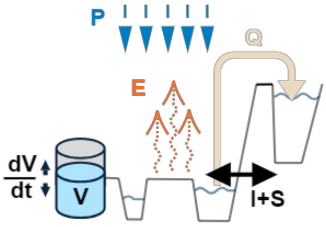


Figure 4.6: Water balance in the polder

The water balance should close on the long-term, because of the conservation of mass. In the study area there have not been major changes in landuse, length of canals or water storage (Delfland, 2022), meaning that there no (big) changes in the storage of the polder, which means that $\frac{\text{dV}}{\text{dt}} = 0$. This results in Equation 4.2.

$$\frac{dV}{dt} = I + S + P - E - Q_{\text{pumped}} \tag{4.1}$$

$$I + S = E + Q_{\text{pumped}} - P \tag{4.2}$$

That means that the sum of the inlet and seepage is equal to the discharge and the evaporation minus the precipitation. The Q_{pumped} is the sum of all the pumps operating in one polder. The inlets has the biggest contribution of the total, since the seepage is small in this area (Delfland, 2022). Consequently, the volume that is let into the polder can be estimated by neglecting the seepage (S) term. If the calculated value for I is positive, it means that water is coming into the polder, while if the value is negative, water flows out of the polder. Another explanation for a negative value could be measurement errors or the error in the precipitation estimation.

The water balance is important to take into account when the performance of a machine learning model is evaluated. The water balance poses a physical constraint, which is the conservation of mass, which cannot be violated. Because a data-driven model cannot take this into account, it is important to consider the water balance afterward. The water balance can help for the reliability of the model and makes it possible to detect any discrepancies in the data.

For the Duifpolder the inlet is calculated as the sum in [mm] for each day, week, month and year and the result is presented in Figure 4.7. In this figure it can be seen that mostly water is let into the Duifpolder in the summer months. Sometimes the value for water inlet is negative, meaning that the amount of precipitation is larger than the amount of water that is being pumped out or evaporated. That could mean that this water is stored in the system, for example refilling the groundwater or seepage to other polders.

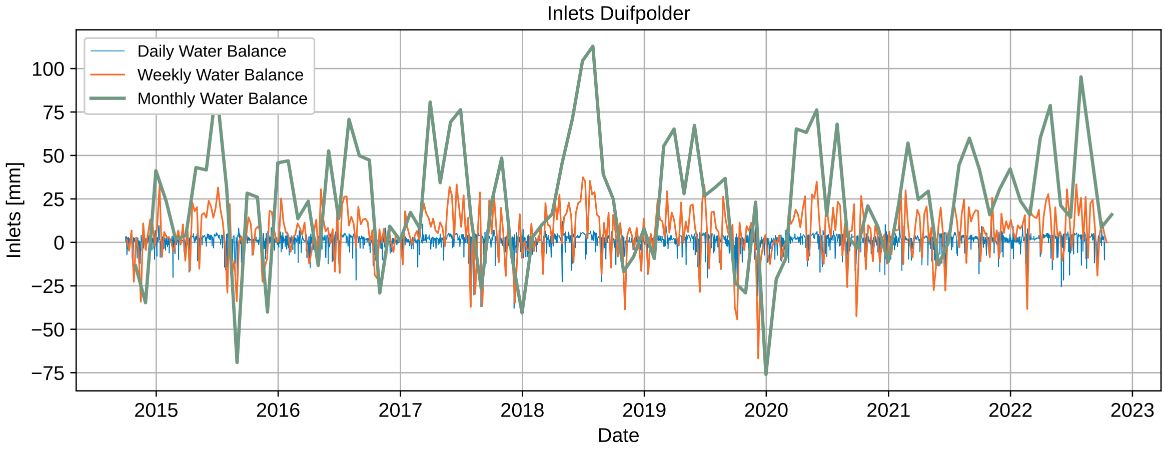


Figure 4.7: Daily, weekly and monthly 'inlets' in the Duifpolder.

The amount of water that is let in, differs over the years. To facilitate a comparison between these different years, a cumulative plot of the 'inlets' has been generated and is presented in Figure 4.8.

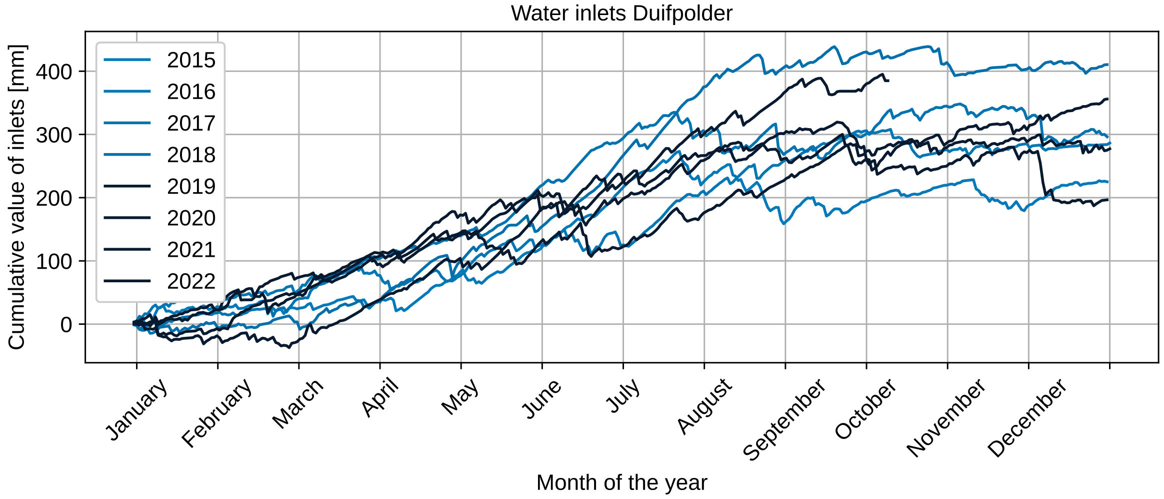


Figure 4.8: Plot of the cumulative amount of the inlets in the Duifpolder for the different years 2015-2021

The water balance for the three polders is calculated and compared. In Figure 4.9 the cumulative value of the inlets for every year is presented. It can be seen that the value for the inlets in 2018 is very high in all polders.

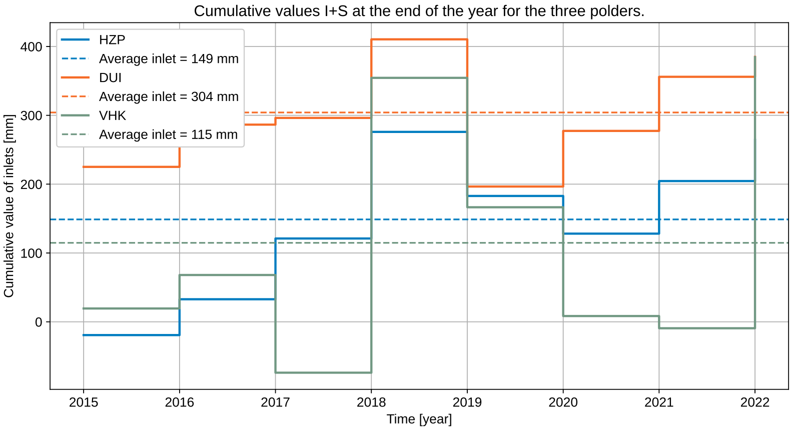


Figure 4.9: Cumulative value of the 'inlets' at the end of each year* in the Duifpolder, Holierhoekse en Zouteveensepolder en de Vlaardingen-Holierhoek polder. The dashed line presents the average of all the years together.

*In this figure the year 2022 means the value for 'I+S' until 10-10-2022 since the data was available until this point.

To check whether estimating the volume of the inlets with this method is suitable, the outcome will be compared to the precipitation deficit that has been calculated for every water authority by De Lange and Koomen (2023). Figure 4.10 shows the precipitation deficit for Delfland for the years 1976, 2018, 2022 and 2023. It can be seen that the two years 2018 and 2022 have a large precipitation deficit. If the precipitation deficit is high, it means that an area receives less precipitation and/or evaporates more than expected over a period of time. This leads to a shortage of water resources, which has an effect on the environment, the agricultural and human activities. For example, the high value for 'inlets' in 2018 can be explained by the fact that 2018 has been a dry year with a long period without precipitation, resulting in a high precipitation deficit.

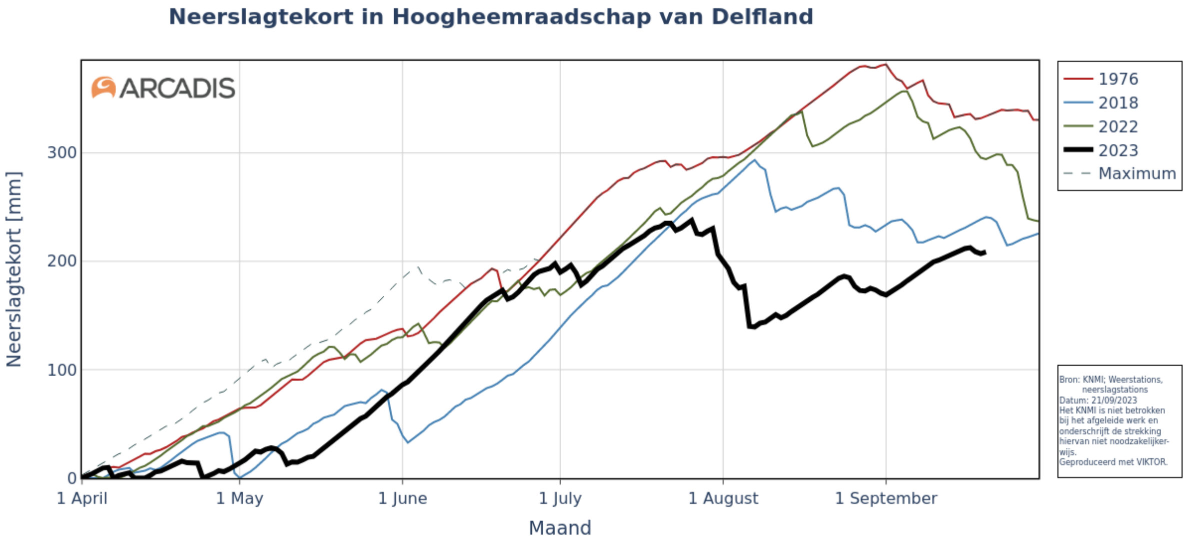


Figure 4.10: Precipitation deficit in Delfland (De Lange & Koomen, 2023)

Using this method gives a rough estimate of the volume of the inlets, and it can be seen that if the precipitation deficit in a year is higher, the calculated volume of the inlets is generally higher as well.

4.2.2. Connection polders

Considering the case study area, the three polders are not connected. The Duifpolder is separated by the boezem canal named the 'Vlaardingse Vaart' and the other two polders are not connected with a canal or culvert. That means that, in principle, the polders are not connected and can be analyzed separately.

4.2.3. Discussion and recommendations

Except for the recommendation of looking into ways how to observe the amount of water entering the polders from the boezem, there are some recommendations that could be done with the current data to get an idea of the volume. If a data-driven method is used, the result is based on the time series of the observed values and there are no other equations or physical constraints used. To avoid outcomes that are physically not possible, it is advised to consider the water balance. The problem when checking the water balance is that there are no observations of the inlets and seepage. These missing observations are especially a drawback during long periods of no rain, when a lot of water is let into the polders. This is a problem that is hard to overcome and results in a underestimation of the total amount of pumped water from the boezem, since the inlets are not taken into account.

Even though the estimate of the inlets cannot be compared to the actual value, it would be interesting to compare the amount of inlet and seepage for the different polders in the area and see if it might be possible to construct a model or an empirical formula that represents this fraction. If an empirical formula is available based on all the polders in the area, this can be used as additional input to the model. Additionally, this empirical formula can be based on the type of polder or other characteristics if the data of also other water authorities is used. This empirical formula could be validated if for a few 'typical' polders the inlets will be measured for a given period of time.

It would be advised to check the water balance of the predicted discharges and compare this value to the value for the inlets of the observed values. In this way it can be seen whether the model over- or underestimates the total amount of water that is let into the polder.

4.3. Feature selection and concentration time

As presented in Appendix A.3 there is a lot of data available. To develop a model with efficient computational performance, an investigation into the key features and lag times is done to avoid unnecessary computational time.

4.3.1. Features random forest model

There are 12 different models created to predict the sum of the discharge for different time horizons in 3 different polders. In this section the features that are used are discussed. Next to that, some research is done on the concentration time of the polder. This can be used to determine the number of hours of the rainfall data that has to be considered for the prediction of the discharges. Lastly, the most important features for the determination of the discharge for the different time horizons are presented and the inputs that will be used for the LSTM model are discussed.

Table 4.1: Pumps of N49 that are selected to predict the discharge

Polder name	Pump number see 3.1	Max. capacity [m^3/min]
Holierhoekse en Zouteveensepolder (HZP)	111102	68.6
Duifpolder (DUI)	106101 (1)	32.9
	106102 (2)	12
Vlaardingen-Holierhoek (VHK)	124101	15

For the random forest model the values at a specific time step of the following features are used:

Water level
Discharge
Radar
Raingauges
Season
Day of the year
TG (temperature)
Q (global radiation)
UG (humidity)
EV24 (reference evapotranspiration Makkink)
Time steps no pumping
Time steps since last rain
Water balance
Prediction rain next 3 hours

4.3.2. Concentration time

In order to know which lagged values and rolling means need to be added to the dataframe for the random forest regression and the size of the input sequence for the LSTM model, it is interesting to look at the time it takes until most of the precipitation ends up in the canals. This time is often referred to as ’concentration time’, meaning the time it takes for water to travel from the most distant point in a watershed to a particular location within that watershed during a rainfall event. Yuswo, 2022. It depends on the polder and the type of land use in that polder. In a more urban polder this time will be much shorter than in a rural polder for example. Other factors that influence this ’concentration time’ is the type of soil, the distance between the polder canals, the presence of ditches, the initial conditions and the intensity of the precipitation.

A short analysis of the concentration time has been done by the earlier mentioned hydrologist of Delfland. A simulation of a linear reservoir for the total polder discharges and the average precipitation of the 10 rain gauge stations has been made using the hourly values to estimate the percentage of the rainfall that has been discharged in every time step. In Figure 4.11 it can be seen that about 90% of the precipitation has been discharged within two days. This graph is an average of all the polders and the volume of water that is discharged in the first hours will be larger in a urbanized polder with a lot of hardened surface.

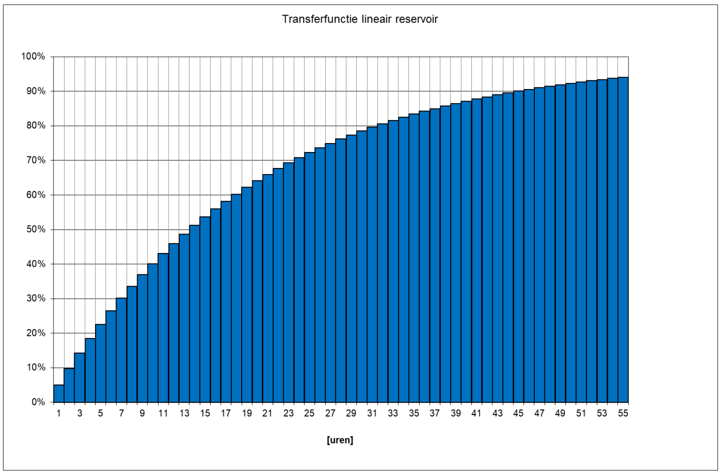


Figure 4.11: Transferfunction linear reservoir for all polder and the average precipitation rain gauge stations

Data Analysis polders in N49

The different lag times of the hourly sum of the rain gauge and the hourly sum of the radar data of the three polders has been compared to the total hourly discharges of the polders. It can be seen that precipitation up to 24 hours ago has the biggest influence on the hourly sum of the discharge. Especially the lagged values of 8-10 hours ago influence the discharge at the current time step.

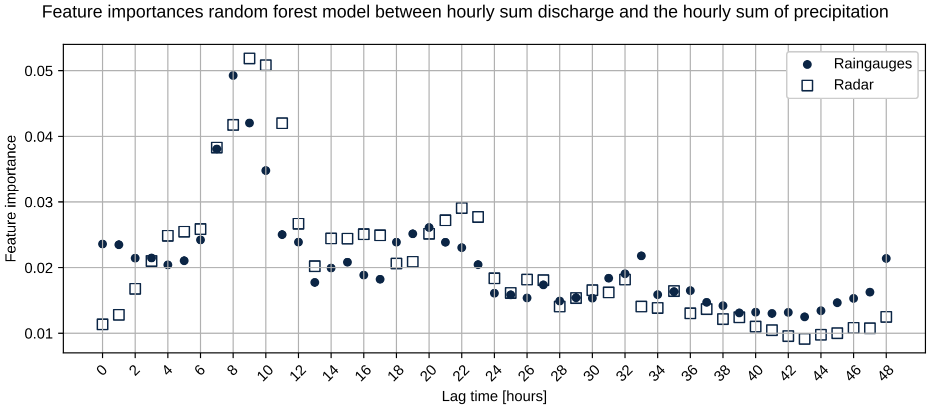


Figure 4.12: Random forest feature importance analysis for the prediction of the discharge given the precipitation on lag times.

4.3.3. Added features

It is important to consider also the current state of the system, which is determined by the observed values of the past hours, days or weeks. For example, the estimation of the total amount of discharge can be different if there has been a lot of rain in the hours before. To take the previous conditions into account, the rolling sum (sum of previous values) and lagged values are added to the dataframe in order to 'capture' the state of the system.

The rolling sum of all the columns are considered as features. It has been chosen to include the sum of the past 1, 2, 8 and 24 hours as additional features to the dataframe (rolling sum of 4, 8, 32 and 96 time steps). This has been done because it is a quickly responding system and the most recent measured values of discharge and water level have the greatest impact on prediction. The changes on the short term will mostly be based on the observed events of the past few hours. The 24 hour time step of the sum of the past day has been included in the features as well, to take also into account some previous events. These added features have a high correlation, meaning that an individual feature will not have as much influence as expected. This causes some 'multicollinearity' in the model Daoud, 2017.

In the study of the autocorrelation and the partial autocorrelation (Section 3.2.2.3) it has been found that the correlation between the observed water levels was very high on the short term. To give the model an idea of the gradient of this variable the lagged values of the water level (previous measured values) of 15 minutes, 1, 2, 4 and 8 hours are added to the dataframe.

4.3.4. Most important features random forest model

In the Figure 4.13 the average relative importance of the most important features are presented. The figure shows the average importances for the three polders together, split into the three time horizons. The higher the 'Importance', the bigger the influence of this feature on the output variable 'Sum discharge next ... hours'.

It can be seen that the importance of other features than water level and discharge increase as the amount of hours that is predicted increases. This can be explained because the current observed value of for example the discharge is relatively more important for the prediction of the discharge on a short time horizon, than for a bigger time horizon. Other features that are important are the rolling sum of the incoming radiation and the evapotranspiration (EV24). These features are correlated, because the value for the evapotranspiration is calculated using this radiation. Evapotranspiration has been selected as a feature in the LSTM model since this is one of the components of the water balance and directly influences the amount of water in the polder.

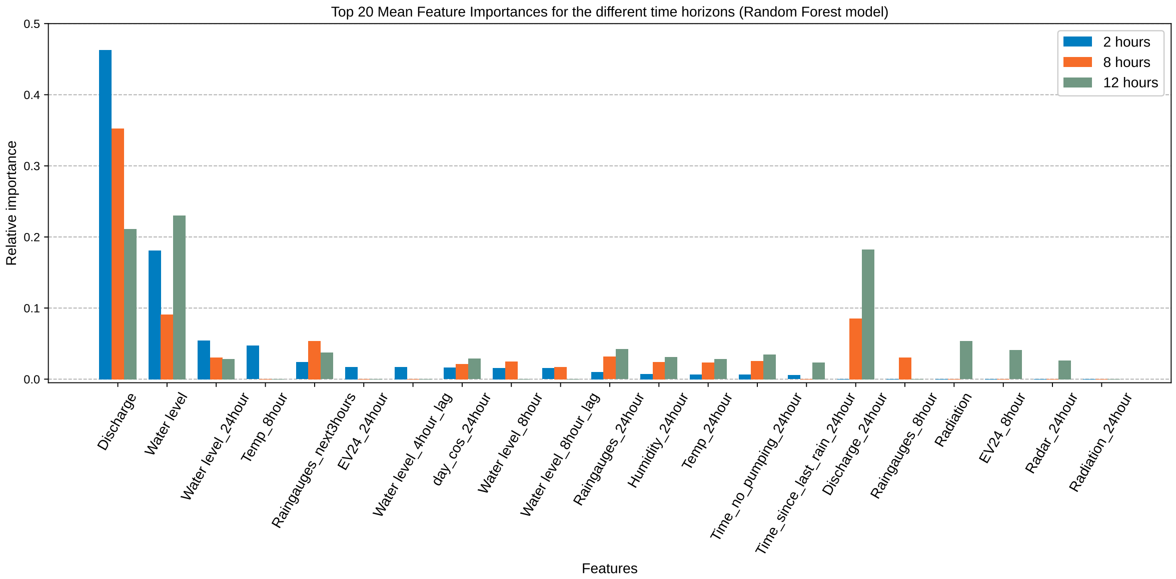


Figure 4.13: Average feature importances based on random forest Regression for the four different pumps depending on the different time horizons.

In Figure 4.14 the features with the highest importance summed for the three different time horizons is presented. It can be seen that the current discharge and water level is the most important for the prediction of the sum of the discharge. Also the rolling sum of the discharge and the water level of the past 24 hours turns out to be an important feature for the prediction of the discharge. The 'prediction' of the precipitation is also influencing the prediction.

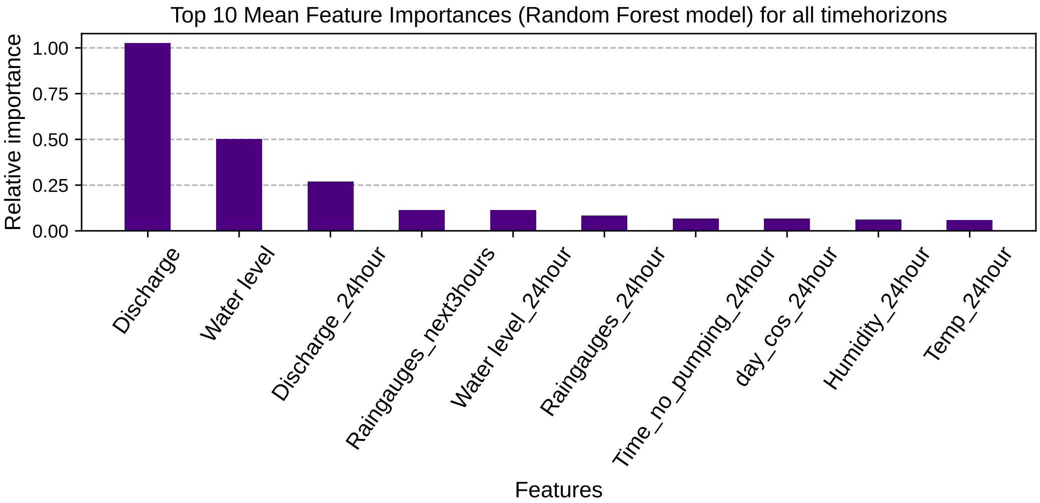


Figure 4.14: Top 10 features based on random forest Regression summed for all time horizons.

Features LSTM model

According to the random forest model, the most important features for the determination of the sum of the discharges are the current discharge and water level. That means that these two variables will be part of the input for the LSTM model. Next to that, it is favorable that the LSTM model reacts to the observed precipitation and the weather conditions. That is why the observations of the raingauges and evapotranspiration are added to the input. Since evapotranspiration and global radiation are correlated with each other (see Figure 4.15), it has been chosen to include only one in the input of the LSTM.

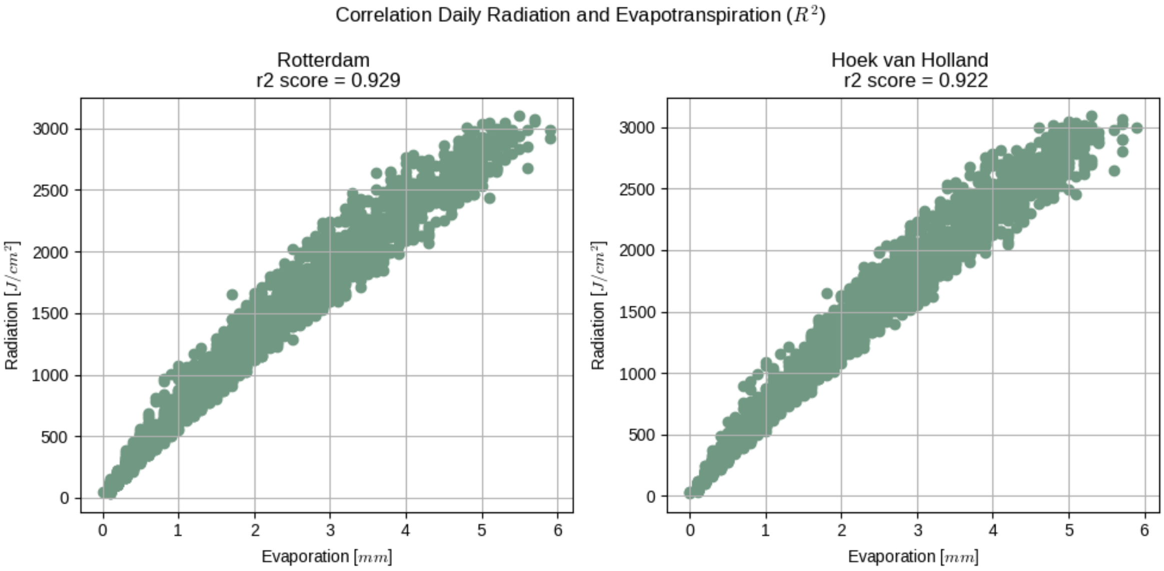


Figure 4.15: Correlation radiation and evapotranspiration

4.3.5. Discussion and recommendations

Additional features

It is possible to further improve the model by adding features that can improve the ability of the model to predict discharges. A few suggestions are:

- Soil moisture content (VanderSat SATDATA-3.0)
- Nowcasting data (Imhoff et al., 2022)

- Using evaporation data with higher temporal and spatial resolution (KNMI data)
- Polder characteristics (land use, area, pump capacity, reference water level, soil type, etc.)

The soil moisture content will provide the model with some information about the state of the model. If the soil is very saturated, the response to a rainfall event will be different than if the soil is very dry. Another feature that could be added is the nowcasted precipitation forecast, which is the prediction of the rainfall up to 6 hours ahead based on extrapolation of recent radar rainfall maps.

The next suggestion is to add KNMI data with a higher temporal and spatial resolution. In this research the daily estimated evaporation data of the station Hoek van Holland is used for the whole area, and the different land-uses in the area are not taken into account. The evaporation depends on the percentage of land that is vegetated and which part is open water, and differs per polder area. It would be better to include evaporation estimations that have a higher spatial and also temporal resolution. The same is advised for the radiation and temperature data obtained from the KNMI.
The polder characteristics can be added to a deep learning model. An example of implementing catchment characteristics in hydrological modelling is presented in the research of Kratzert et al., n.d. Adding these polder characteristics will result in possibilities for using pre-trained models for a new pump. This can be useful because for a newly installed pump since no data is available on which a model can be trained. More features that could be added to the model are discussed in Appendix A.3.

Data quality and availability

The results of the prediction of the discharges and the water levels using data-driven models like a RF regression and a LSTM model, heavily depend on the quality and the availability of the input data. If this data is not available, this will result in problems for the model performance, since this is the only input of the model. It should be considered what would happen if for example at one location the measurement device is not working anymore, since the current measurements are essential for the two models.

Difference importance radar and rain gauge data

The feature analysis shows that the observations of the rain gauge data have more influence on the sum of the discharge than the radar data. The correlation coefficient of the daily sum of the discharge is compared to the sum of the precipitation, based on radar and rain gauges separately. The correlation of the rain gauge data with the discharge is slightly higher. This could be due to that rain gauges data are direct measurements and represent the rainfall on a specific location. The estimation of the rainfall based on radar data is influenced by atmospheric conditions and terrain properties, which might lower the correlation. The difference in correlation between the variables can also be influenced by errors or noise.

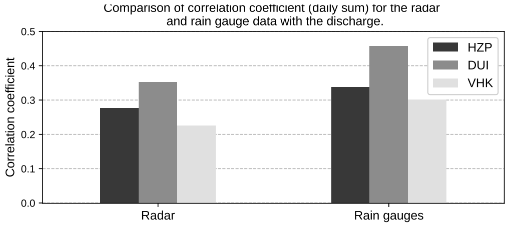


Figure 4.16: Comparison between the correlation coefficients for between the radar and rain gauge data and discharge for every polder in N49.

The correlation between the sum of the hourly radar and rain gauge data is lower than 0.8 in all polders, see Figure 4.16. This means that more study needs to be done on which data of the precipitation is more representative for the area.

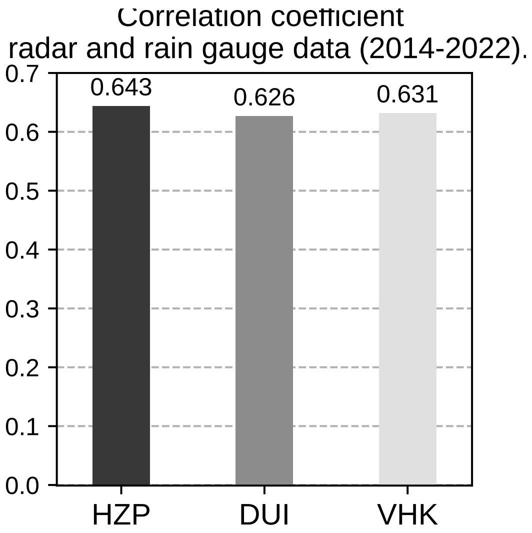


Figure 4.17: Comparison between correlation coefficient radar and rain gauge data in every polder in N49.

Sparse variables

The effect of precipitation on the predicted discharge was not as high as one would expect from a hydrological point of view. This can be explained by the large amount of zeros in these timeseries, so-called ‘sparse data’, which makes the influence on the pumped discharge smaller. Sparse data refers to datasets in which most of the elements are zero or empty (Greenland et al., 2016). It would be recommended to take this sparse data into account for the construction of models in the future.

Selection of rain gauge station

There are 10 rain gauge stations in the area, measuring precipitation every 15 minutes. In the current model, each polder is coupled to the rain gauge station that is located nearest to the pump that is considered. This is a very simple method that assumes that the precipitation in the whole polder is exactly the same at the observed value at the closest rain gauge. In order to increase the spatial resolution of these observations and thus the estimation of the rainfall using these rain gauges, this can be done with two commonly used methods; Kriging or Inverse Distance Weighting (IDW). Kriging is a geostatistical method that estimates values at unobserved locations by modeling spatial autocorrelation in the data (Setianto & Triandini, 2013). It takes into account not only the distance between points but also the spatial structure of the data, making it particularly effective for spatially correlated data. In the Inverse Distance Weighting (IDW) method, it is assumed that the strength of correlations and similarities between neighboring points is directly related to the distance between them. This relationship is defined as a reverse function of the distance from each point to its neighbors. However, it is crucial to note that determining the radius of what constitutes a ’neighbor’ and the specific exponent in the distance function are critical considerations in applying this method (Setianto & Triandini, 2013).

Availability forecast of precipitation data

In the current model only the future observed values for the precipitation are used as ‘forecasts’ it would be advised to use real forecasts and to also add forecasts of other features. Because the historical predictions were not available, the shifted observed values of the rain gauge data are used as a simulation of the ‘forecast’ for the precipitation of the next 3 hours. These ‘predictions’ are far more accurate than currently possible. To make this forecast more realistic, for example a random noise can be added to these observed values.

Availability of final product of the radar data

Another important point to take into account when interpreting the results is that it takes about one hour before the raw radar data is available. These data are then post-processed in an early reanalysis by combining data with measurements. The early reanalysis product is available after 1-2 days. For the final product it can take up to 60 days before this is available. This means that in reality the ‘first product’ will be used instead of the ‘final product’ of the radar data, which is likely to be less accurate.

4.4. Optimization of the hyper parameters of the Random Forest model and the LSTM model

In this section an overview of the different hyperparametersets that have been tried for the random forest models are presented. Next, the optimization process of the LSTM model is visualized. The optimal hyperparameters are searched using the Optuna algorithm for the random forest model as well as the LSTM model.

4.4.1. Random forest model

For the random forest model the hyperparameter `n_estimators`, `max_depth`, `min_samples_split`, `min_samples_leaf` have been optimized using the Optuna algorithm, discussed in 3.3.3. In Figure 4.18 can be seen that there is a wide variation in the chosen combinations of hyperparameters for the different trails.

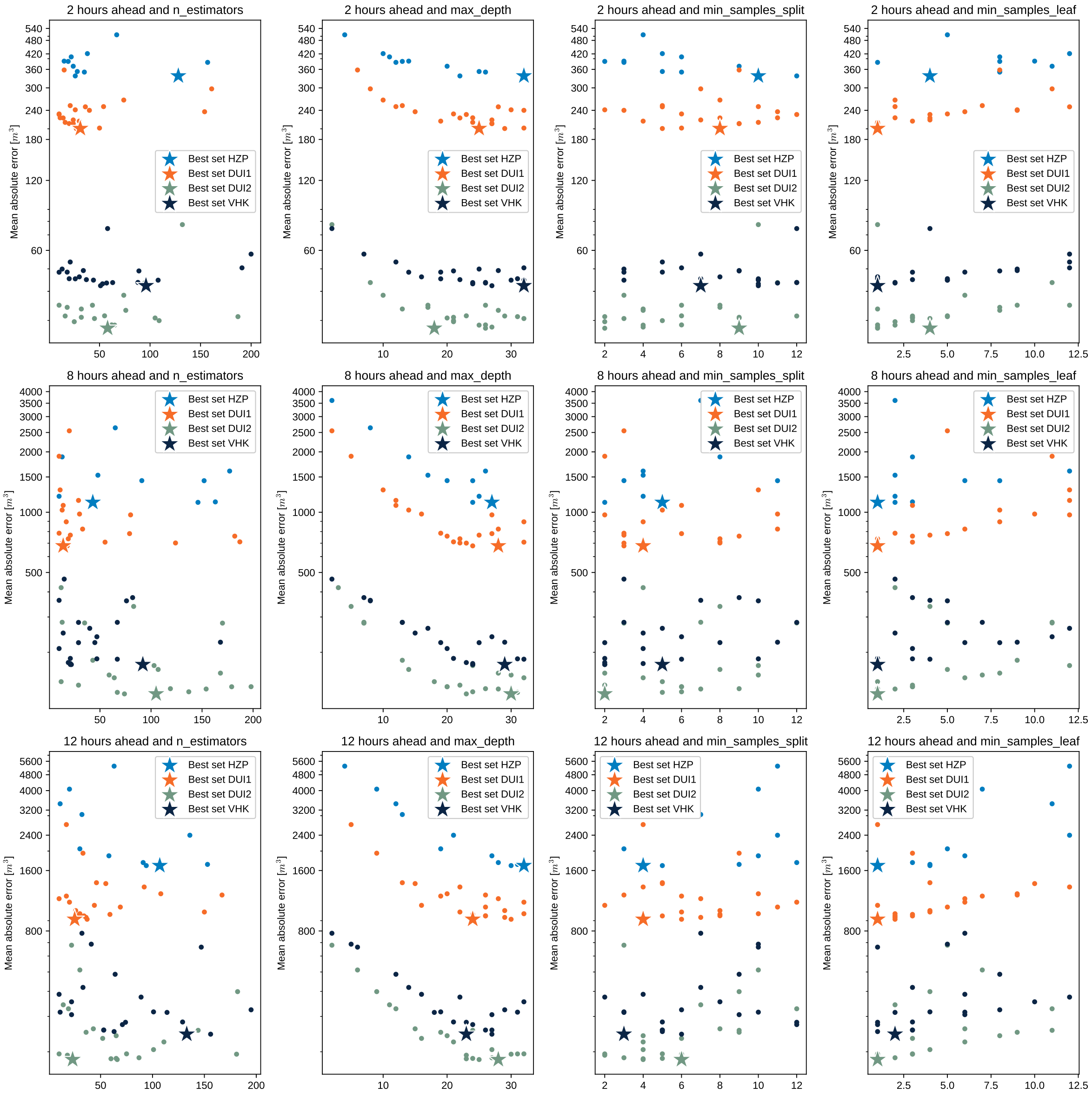


Figure 4.18: Hyperparameters of the trails for the optimization of the RF model.

The y-axis presents on a log-scale the Mean Absolute Error of a certain hyper parameter set. That means that for example the optimal set of the HZP polder has a MAE of about 338 m^3, and a value of 128 for the `n_estimators`, a value of 32 for the `max_depth`, 10 for the `min_samples_split` and 4 for min_samples_leaf.

For n_estimators (the number of estimators) the optimal sets are mostly around 50 and sometimes a bit higher between 100 and 160. The hyperparameter max_depth is very similar for the different polders and is mostly between 20-30. The optimal values for the min_samples_split are differing a lot for the different time horizons. The optimal hyperparameter set has a value for min_samples_leaf below 5 for all the pumps and timehorizons.

It can be seen that the hyperparameter max_depth has the biggest influence on the performance of the model since there is a clear trend observable in the MAE of the different trails. This means that it could be considered to narrow down the search space for this hyperparameter if more models are created and optimized. The trend of the variable min_samples_leaf is not as clear as for the max_depth, but it can be considered to narrow down the search space for this parameter as well. This will reduce the computational time of the model.

4.4.2. LSTM

For the LSTM model the training process will be visualized instead of the optimization of the hyperparameters. The optimization for the LSTM model is done in a similar way as the RF model, with Optuna, but of course with different hyperparameters. The LSTM model has been made only for the sum of the discharge in the next 12 hours. The training process of the LSTM model is visualized in Figure 4.19.

Visualization training process

In Figure 4.19 the training process of the optimized model is visualized. It can be seen that the loss decreases over epochs, which means that the model is learning over time.

The best model is trained by evaluating the MSE. The result of the training process is plotted by squaring the outcomes to obtain the RMSE and then scaling this array back to the actual ranges of the observed discharges. The error that is decreasing over the epochs is the error between the model prediction of the discharge in the next 12 hours compared to the observed discharge sum. The dashed line in the figure marks the early stopping of the training process.

The time each epoch takes depends on many factors, for example the number of years that are used for training, the way the model is optimized and the type of hardware that is used. The optimization process for the LSTM models take 1.5 days per pump if in total 6 years of data is used as training and validation data.

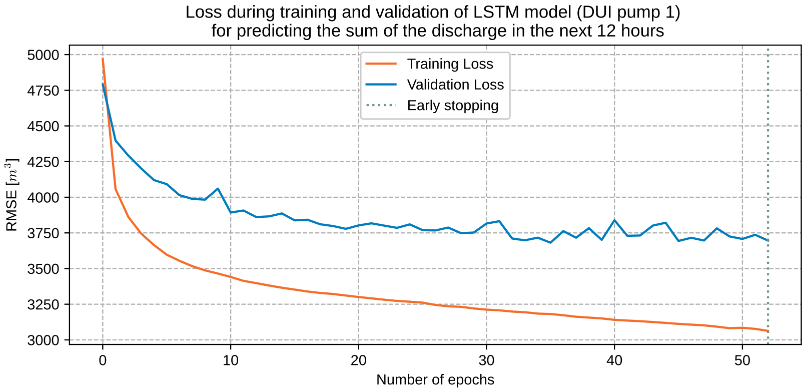


Figure 4.19: Training and validation loss (RMSE) over the epochs

4.4.3. Discussion and recommendations

It would be advised to compare the performance of both the optimized as the standard models in order to see the change in performance. Another idea is to predict the sum of the discharges with a dataframe with an hourly temporal resolution, which will be a rougher estimate, but could be used if the time horizon of the prediction is bigger.

Random forest model

For the optimization of the random forest model, it would be interesting to compare the performance of the model using k-fold cross-validation and using the whole dataset at once to search for the best hyperparameters. Another recommendation is to constrain the number of features in the model, what has not been done in this optimization. The hyperparameter search has been done for 20 trails, while a higher number of trails (50-100) will improve the model even more, since for some of the hyperparameters the convergence towards an optimal value are not clear.

LSTM

The LSTM model could be improved by adding more features, as for example the predicted precipitation. Another advice is to perform the hyperparameter optimization based on the different k-folds, as has been done for the RF model. Another recommendation would be to add the discharge prediction for the next 2 and 8 hours, and to see how the LSTM model predicts these discharges compared to the benchmark models and the RF model.

Another suggestion is to try different types of LSTM networks, for example a 'deep' LSTM network, containing multiple LSTM layers. Another option is to combine LSTM with a Convolutional Neural Network (LSTM-CNN) (X. Li et al., 2022). Another new type of LSTM is combining Step-sequence framework, which showed great potential in predicting the daily rainfall and multiple step- ahead discharge predictions in Yin et al., 2022.

4.5. Performance of the Random Forest and LSTM model

For the four pumps presented in Tables 4.2 and 4.3 the sum of the discharge for the upcoming 2 and 8 and 12 hours is predicted with a random forest model. This sum is compared to the observed sum of the discharge in that timeframe and the two benchmark models (ReRengAI and naïve). A LSTM model is made for the prediction of the 12 hourly sum for the four pumps and also evaluated by computing the RMSE in Table 4.4.

4.5.1. Comparison performance per polder pump

In this section the error of the different pumping stations is evaluated.

Prediction 2 hourly sum

The prediction of the sum of the discharge in the next 2 hours is close to the actual sum of the next 2 hours as can be seen in Figure 4.20. A summary of the model results for the predictions of the discharges in the next 2 hours is given in Table 4.2. It can be seen that the error of the random forest model is always lower than the ReRengAI model. The error of the pumping station of the HZP is lowest for the naïve model.

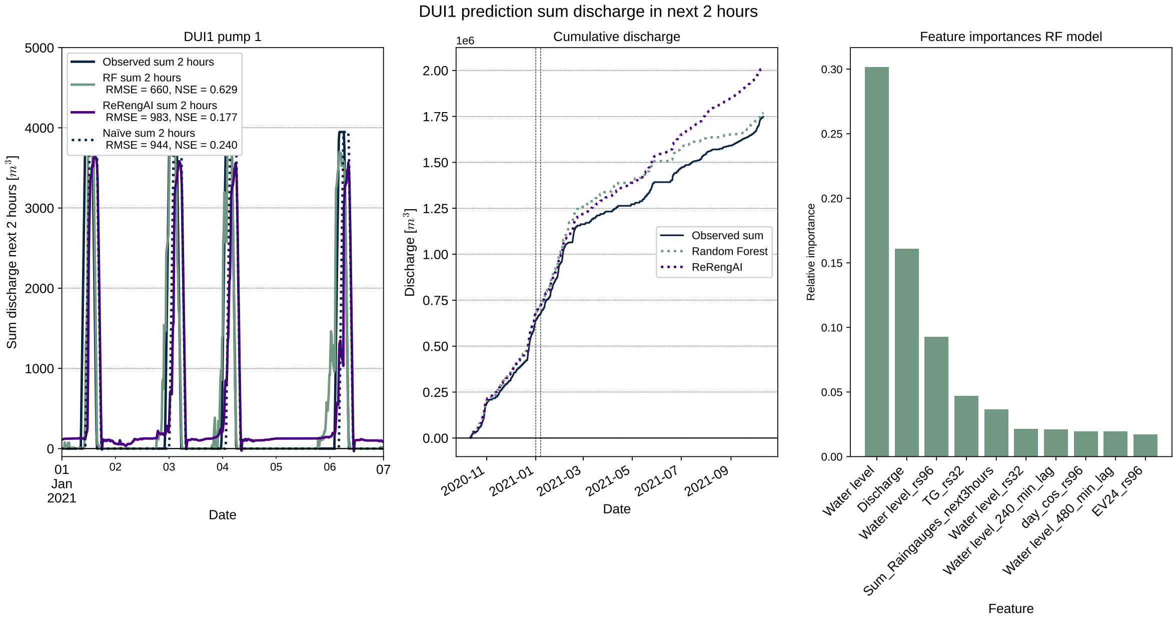


Figure 4.20: Comparison random forest, ReRengAI and naïve model for pump 1 of the Duifpolder (DUI1)

Table 4.2: Comparison RMSE models discharge 2 hours

sum 2 hours [m^3]	Random forest	ReRengAI	Naïve
HZP	1046	1090	**851**
DUI1	**660**	983	944
DUI2	**81**	126	112
VHK	**83**	108	102

Prediction 8 hourly sum

The model results for the predictions of the discharges in the next 8 hours are given in Table 4.2. It can be seen that also for the predictions of the sum of the discharges in the next 8 hours the error of the RF model is smaller than for the ReRengAI model, except for the HZP polder, in which the naïve model performs has the lowest RMSE.

Table 4.3: Comparison models discharge 8 hours

sum 8 hours [m^3]	Random forest	ReRengAI	Naïve
HZP	5781	6619	**5631**
DUI1	**2791**	3460	3983
DUI2	**480**	646	677
VHK	**611**	723	662

Prediction 12 hourly sum

The prediction of the sum of the discharge in the next 12 hours for the two pumps in the Duifpolder is presented in Appendix A.5. The random forest model outperforms the ReRengAI model in all polders as presented in Table 4.4. In this table also the RMSE of the LSTM and naïve model are presented. In which the LSTM model is better in estimating the discharge in the HZP and VHK polder.

Table 4.4: Comparison models discharge 12 hours

sum 12 hours [m^3]	Random forest	ReRengAI	LSTM	Naïve
HZP	9134	10748	**8098**	9555
DUI1	**3714**	4277	3852	5087
DUI2	**748**	906	763	975
VHK	1047	1273	**845**	1161

4.5.2. Comparison N49

To compare the models to the Sobek RR model, the predicted values for the four pumps are summed and compared to the outcome of the Sobek RR model. This is done by converting the predicted sums back to observed values per hour.

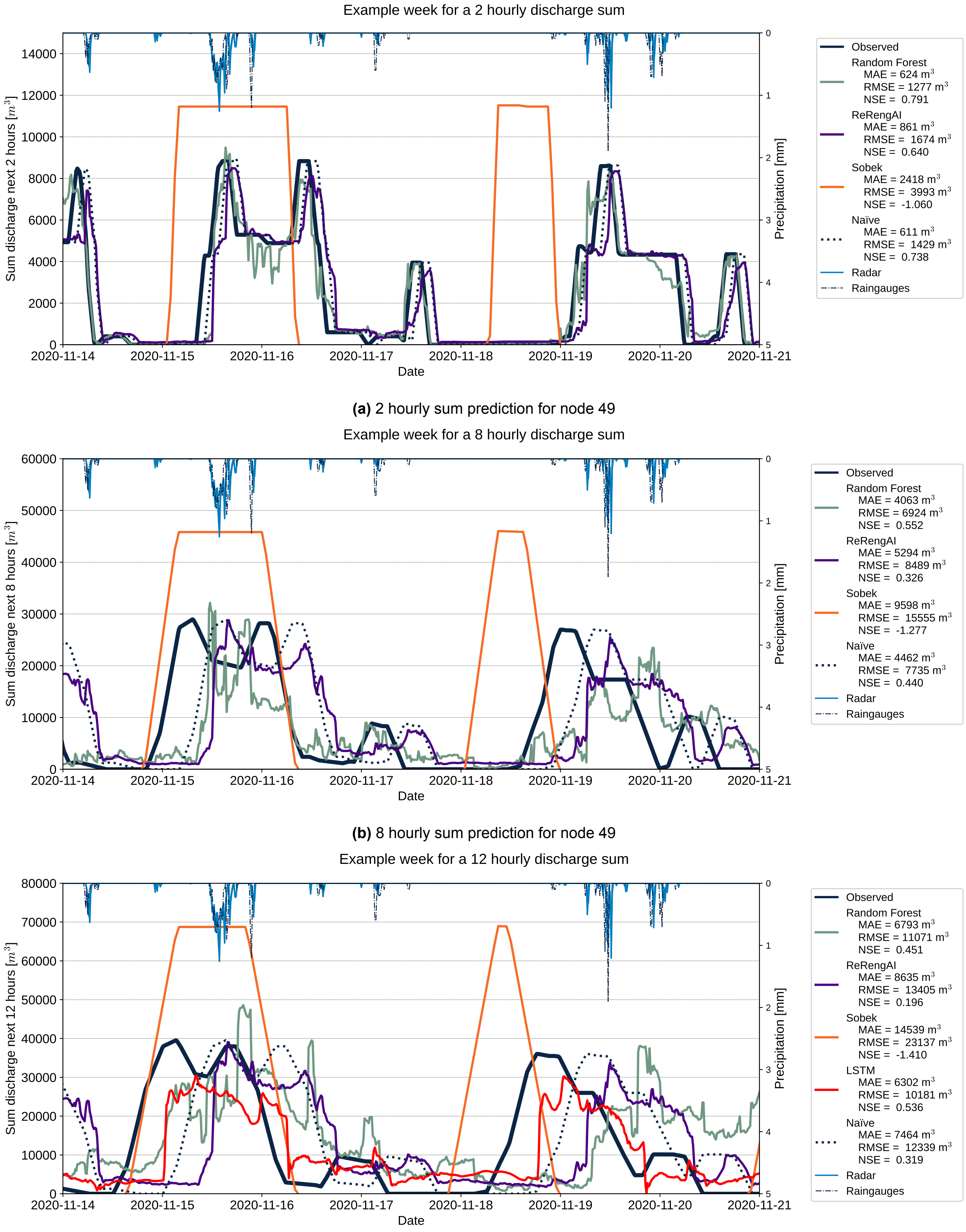
RMSE

In Table 4.5 the RMSE of the different models is presented for the gradient of the discharge. This means the predicted and observed value is evaluated at each time stamp and the average RMSE is calculated for the whole test year. It can be seen that the RF model performs the best, except for the discharge predictions of the 12 hourly sum. The 12-hourly sum is predicted with a LSTM model and this model outperforms the RF model.

Table 4.5: RMSE of the observed and predicted values for the three different models

	RMSE [m^3]				
	Random forest	ReRengAI	Sobek RR	Naïve	LSTM
2 hours	**1,277**	1,674	2,418	1,429	
8 hours	**6,924**	8,489	15,555	7,735	
12 hours	11,071	13,405	23,137	12,339	**10,181**

A graph of the gradient of the different models of the second week in November 2020 is presented in Figure 4.21. This is done for the three different time horizons. It can be seen that based on the predictions in the test year 2021 the random forest model performs best for the 2 and 8 hourly discharge prediction in terms of RMSE and NSE. The worst performing model is the Sobek RR model. The Sobek model is lumped and is only predicting that either all pumps are on or off, as discussed in the introduction. It can be seen from the figure that the Sobek model often overestimates the actual discharge by assuming that all the pumps are on at the same time. In reality, it could be that only one pump is operating instead of all the pumps in the node, which gives a wrong estimation of the discharge.



(a) 2 hourly sum prediction for node 49

(b) 8 hourly sum prediction for node 49

(c) 12 hourly sum prediction for node 49

Figure 4.21: Comparison models total discharge node 49 zoomed in (HZP, DUI1, DUI2 and VHK) for the prediction of the sum of the next 2, 8 and 12 hours.

Cumulative value

To get an idea whether the models are over- or underestimating the discharge, a cumulative plot of the RF, ReRengAI and LSTM model is made. In this part the sum of the discharges of all the predictions is calculated. This is done by summing the discharges from all the four pumps. The RF model performs better than the ReRengAI model for the 2 hourly discharge sum (see Figure 4.22).

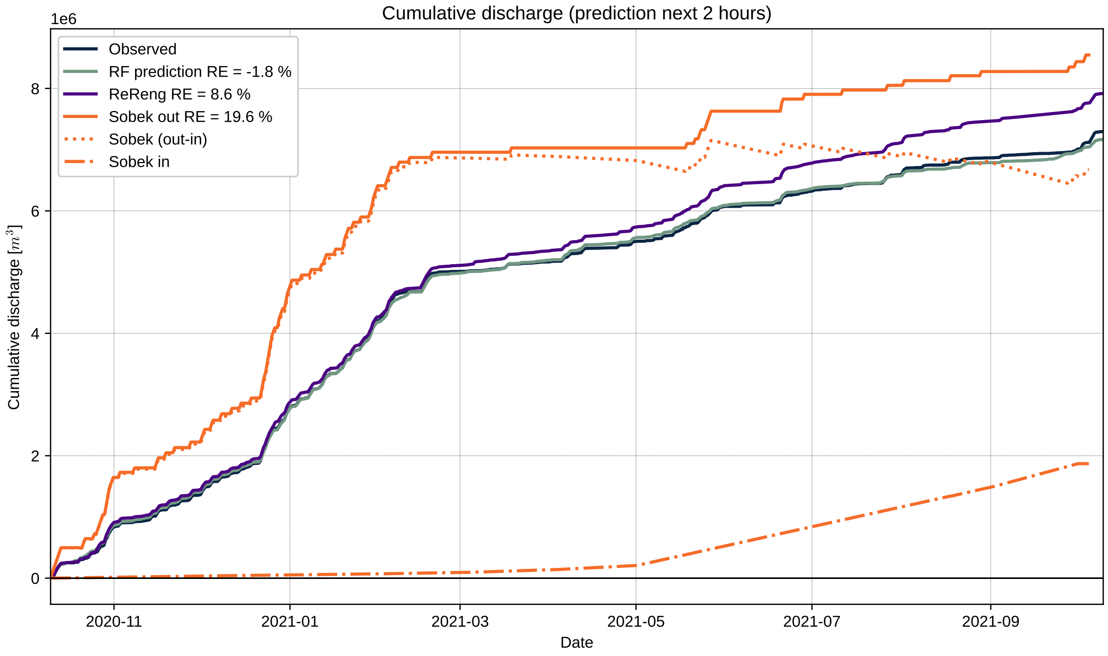


Figure 4.22: Cumulative plot (2 hours ahead) of the observed discharge of node 49 (HZP, DUI1, DUI2 and VHK) in year 2021. The predicted discharge of the RF, ReRengAI and the Sobek model is plotted and compared with the actual measured values.

For the 12 hour ahead prediction, it can be seen in Figure 4.23 the LSTM model is very close to the end value of the cumulative discharge (error of 0.5%) and is following the observed values quite well.

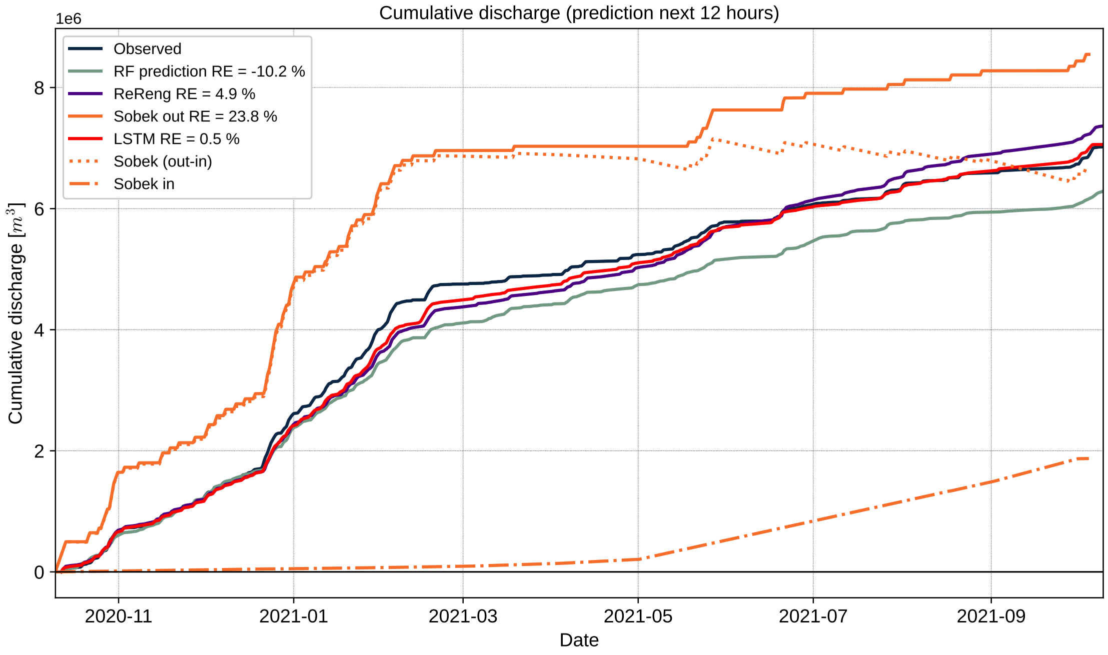


Figure 4.23: Cumulative plot (12 hours ahead) of the observed discharge of node 49 (HZP, DUI1, DUI2 and VHK) in year 2021. The predicted discharge of the RF, ReRengAI, Sobek RR and the LSTM model is plotted and compared with the actual measured values.

The random forest model is the best in predicting the 2 hour sum of the discharge, while for the

12 hour ahead discharge, the best model is the LSTM model, as also presented in Table 4.6. This evaluation is based on the final value for the cumulative sum of the models and does not give information about how well the model is performing at the different time steps.

Table 4.6: Difference observed and predicted cumulative value for the three different models

	Relative error [%]			
	Random forest	ReRengAI	Sobek RR	LSTM
2 hours	**- 1.8**	8.6	19.6	
12 hours	-10.2	4.9	23.8	**0.5**

4.5.3. LSTM model

The LSTM model is used to predict the 12 hourly sum of the discharges of the pump. In Figure A.17 some input sequences and outputs of the two different models are presented in Appendix A.5.3.

To evaluate the performance of the LSTM model for the prediction of the discharge, the predicted value is compared to the observed discharge. For both the LSTM as the naïve model the difference between the observed value and the predicted value is calculated. It can be seen that in general the LSTM model underestimates the sum of the next 12 hours. In the Figure the right (orange) plot shows a big proportion of the errors at a value of 0, meaning that the model predicts exactly the right amount most of the time. Though if this is not the case, the model also often over- or underestimates the sum of the discharge.

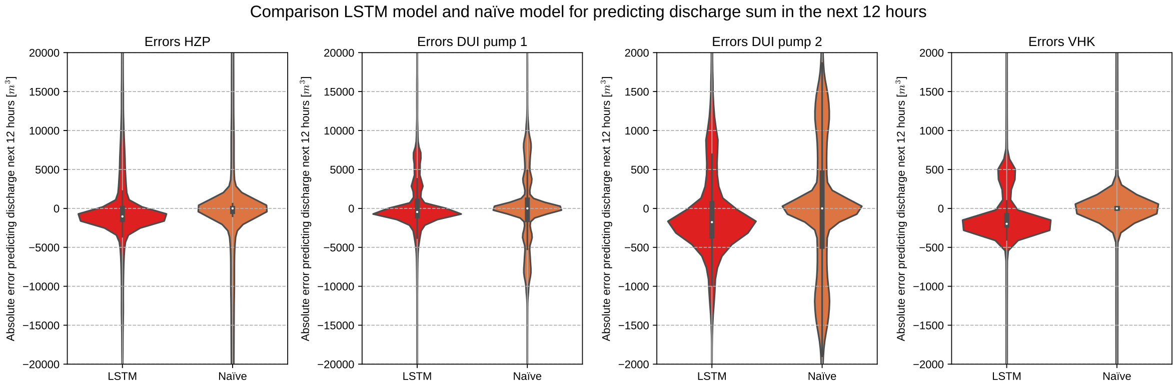
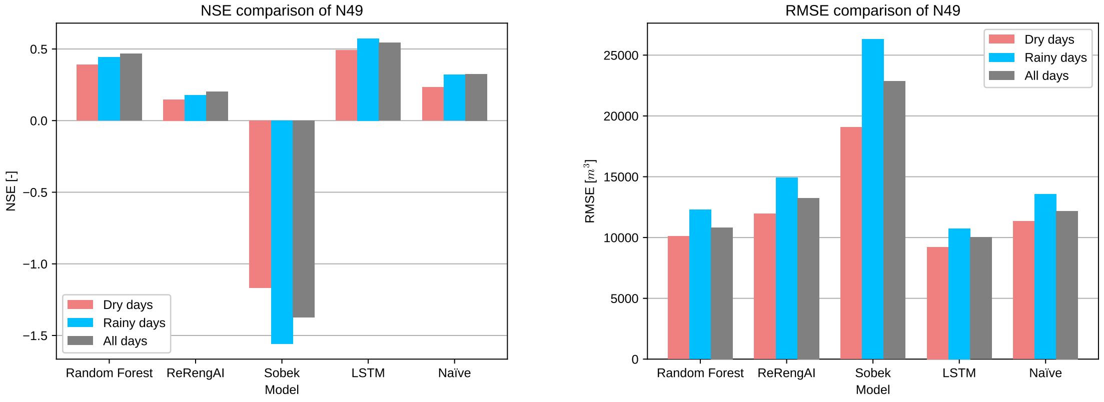


Figure 4.24: Violin plots of the errors of all the predictions made with the two different models

4.5.4. Performance of the models during wet and dry days

The RMSE and the NSE is evaluated looking at the wet and the dry days separately for the prediction of the 12 hourly discharge. The dataframe is resampled to hourly means in order to compare the predicted values with the benchmark models. A wet day is defined if the value for the rolling mean (over a day) of the sum of the radar data of all the three polders is not zero.



(a) NSE of N49

(b) RMSE of N49

Figure 4.25: Comparison NSE and the errors of the different models during the rainy and dry days.

It can be seen in Figure 4.25 that the error during rainy days is generally higher than on dry days. This was expected since during rainy days the pumped discharges are higher and thus the errors are more likely to be high as well. Though it can be seen that for the Sobek model the difference in performance is bigger than for the other models. So during the rainy days the prediction of the Sobek model is worse than during dry days, while the rainy days are crucial for the operation of the 'boezem' pumps.

4.5.5. Discussion and recommendations

The results should be interpreted while keeping in mind the limitations of the research. As discussed in the previous subquestion, there are some ways to improve the overall performance of the RF model as well as the LSTM model. Next to that there are some important considerations to take into account when a data-driven model is used and if the benchmark models are compared.

Data-distribution

For all types of models, the quality of the input data and the amount of available data influence the model performance. However, the data-driven models fully rely on data, meaning that if the data is differently distributed than the data on which the model is trained, the performance will be influenced more than in a conceptual or physical model.

It is important that the training, validation and test data have a similar distribution, as this influences the model performance. That is why the splits in the data are done exactly after a year, since this makes the probability that the data is distributed the same, higher, because the response and water level distribution change through summer and winter. Another idea would be to make a separate model for the response in winter and summer.

Even though the change in summer and winter water level is taken into account, it is still possible that the distribution of, for example, the water levels is not the same for the different years. An example of a change in the distribution of the observed data of the Duifpolder was observed in year 2022 from mid-July, as can be seen in Figure 4.26.

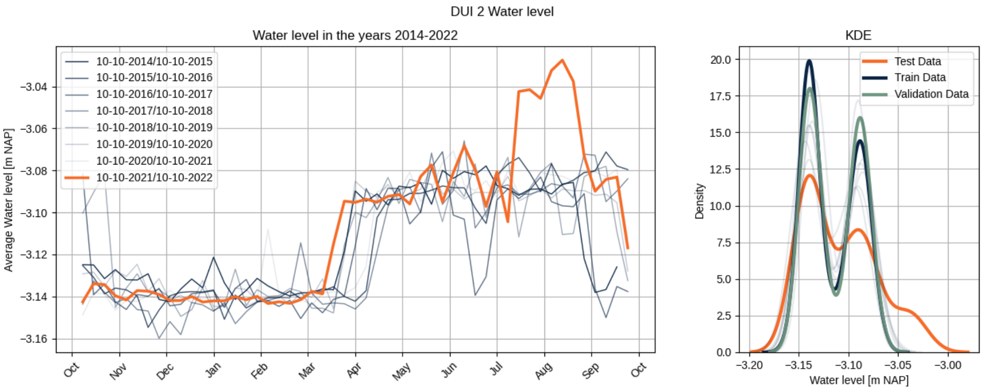


Figure 4.26: The water level and the distribution of the water levels in the training, validation and test years.

This increase in water levels can be explained by the fact that this was a dry year in which the operators tried to avoid a water shortage in the polder. They locally increased the water level at which the pumps are switched on and off, to avoid this water shortage. The result of this adjustment was a bad performance of the random forest model from this time step onward, shown in Figure 4.27.

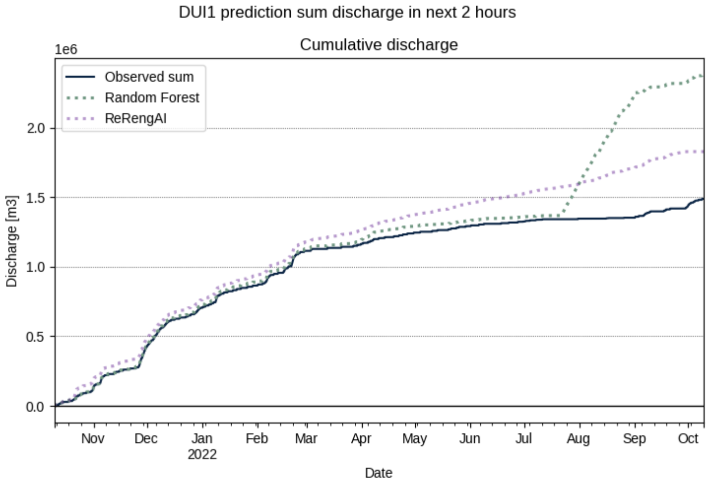


Figure 4.27: The effect of the different data distribution on the model performance at pumping station 1 in the Duifpolder.

The data for the year 2022 has been excluded from the dataset, because the water levels in the Duifpolder (pump 1 and 2) and in the Holierhoekse en Zouteveensepolder (HZP) showed a different distribution compared to the training and validation years. This altered distribution adversely affected the accuracy of the model's predictions. That is why the data from 10-10-2014 until 10-10-2019 have been used as training data and the next year 10-10-2019/10-10-2020 as validation data. In this way, the year 10-10-2020/10-10-2021 is used to evaluate the models (test data).

Comparison performance

Root mean squared error The evaluation is mainly based on the root mean squared error, which does not provide information on whether the model is over- or underestimating the pumped discharges. It only provides information about the magnitude of the average error of all the different time steps in the test year. A recommendation would be to look at the proportion of water coming from N49 in relation to the other Sobek nodes, and then estimate this error for the whole area and look at the effect of the improvement of the prediction for the DSS.

Evaluation of the cumulative value The models are evaluated by looking at the difference between the modelled and the observed values, The difference is expressed in the MAE and the RMSE. The random forest model demonstrated very good performance, slightly surpassing the in-development machine learning model of Delfland (ReRengAI). Additionally, the random forest model outperformed the Sobek model. The Sobek model tended to overestimate discharge in the initial months. The random forest prediction of the cumulative discharge through the year accurately follows the observed cumulative discharge for the 2, 8 and 12 hour ahead prediction, but the ReRengAI model was slightly better for the 8 and 12 hourly sum. If the cumulative value at the end of the test year is considered, the ReRengAI model performed very well. The consideration that has to be made is whether this final cumulative value is more important, or all the predictions separately. The final number of the cumulative sum of the discharge is not very representative for the evaluation of the model since it does not contain any information about the performance of the model through the year. It could be possible that the model is performing very bad at the beginning by underestimating the discharge, while this is compensated in the second part of the year, resulting in an accurate final value, but a very bad actual model performance through the year. This means that the evaluation based on the cumulative value gives an idea of the overall volume that the model estimates, but not if the model is correctly following the dynamics.

Benchmark models

The benchmark models that are used have some limitations that should be considered when making a reliable comparison. First of all the machine learning model of Delfland, ReRengAI, is discussed. This model (further discussed in Appendix A.2.3), predicts the discharges at the given moments in the future instead of the sum. To compare the results, this gradient is converted to the integral of the discharge in the next 2, 8 and 12 hours. It has to be noted that the results are obtained without the split in a training and a test set. This means that the model is likely to overfit to the data and will be lacking in performance for unseen data. In this case it means that the performance of this model is better than it would be in reality, since the model has been trained on this data and could memorize this.

Secondly, it is important to note that the Sobek RR model is not updated every 15 minutes, as the other models are. The model is run at once and is not updated with the 'current state'. This lowers the performance of the model substantially.

LSTM

The results of the LSTM model are not as optimal as they could have been. It can be seen in the distribution of the errors that the mean of the errors differs a lot from the median, which is not a good sign for the performance of the model. This could be caused for example by that the training data was not representative, or that the architecture was not ideal or for example that predicting the sum of the discharge is not a suitable output for a LSTM model. Another explanation could be that the selected features are not the most representative for the prediction of the discharge sum, since they are based on the features of the random forest model. On top of that, the performance of the LSTM model could be improved if for example a combination of a physically-based model and a deep learning model is used, further specified in subquestion 1, modelling approaches (4.1.3.2).

Naïve model

The naïve model, which used the past 2, 8 and 12 hours for the prediction of the next, showed good results as well. This might be because the system is in many cases not changing a lot over a few days, this means the pumped discharges are predictable and could also be estimated with a simpler model. It is recommended to study the performance of simpler models and to see what the added performance is of using more complex models for modelling the discharges. Based on this study, more insight can be obtained on the trade-off between the complexity and the explainability of the used model.

4.6. Implementation of a ML model for the operational control of the water in Delfland

It is important that a model can provide good estimates of the actual discharge in the next hours, but on top of that, there are several important factors that should be considered as well if one wants to apply a ML model in practice. The outcomes of the model are an important input for the DSS model that controls the water levels in the boezem canals and safeguards the area in the daily situation. That is why the objective is to find out more about the requirements for a new type of model to be implemented in the water authority of Delfland. Next to that, some research is done on the general view upon these type of models in the organization. In this chapter, the drawbacks and limitations of a data-driven model and the findings of the interviews with the representatives of the organization are discussed, followed by a discussion and some recommendations.

4.6.1. Drawbacks and limitations data-driven models

The main drawbacks of data-driven models in comparison to physically based models are: the explainability, the interpretability and the fact that there are no physical constraints. Next to that, data-driven models are generally having difficulties with generalizing to new or unseen scenarios that are different from the training data.

Interpretability

The interpretability refers to the degree to which a model's predictions can be understood and explained by humans and focusses on the model's inner workings. The interpretability of a model is important and is easier in the case of a physically-based model, since the predictions are made based on physical laws and equations, which can be interpreted by the water operators. It is more difficult to obtain insight in the process of making predictions for a data-driven model, which has been trained only on the data and has found relations between the input and output.

Explainability

The explainability of a model is very important for the operators to trust the predictions that are made and involves how the outcomes specifically can be understood and justified by humans. The more complex a model is, the less explainable the model is. Though if for example a linear regression model is considered, the explainability is higher, but it will generally suffer from low performance. That means, there is a trade-off in data-driven models between the explainability and the performance (Fu et al., 2022).

Robustness

It is important to consider the robustness of the model to make sure that the model also produces trustworthy results in case of noisy or incomplete data. The robustness can be qualified in terms of consistency and predictability. To increase the robustness of a model, and thus the confidence in a model, the data that is used should be carefully pre-proccessed and validated or the output space of the model should be controlled by imposing constraints in the model. To include these analyses, the system needs to be well understood (Razavi, 2021). Generally speaking, a physically based model has a higher robustness (Setianto & Triandini, 2013). However, the strong robustness of LSTM-Runoff demonstrates the potential of such ML approaches for modeling under changing conditions as well (Dutra & Orth, 2020).

4.6.2. Interviewees

The people that have been interviewed have been anonymized. The interviewees come from different backgrounds, some are more involved in the operation of the models, some have a more governmental role in the organization and others are working more on the internal processes within the organization. This means that the focus in the interviews was slightly different. The complete interviews can be found in A.6 and in this section only some key points are discussed.

Findings

The biggest disadvantages of deep learning is the dependency on the quality of the input data. This influences the performance of the current model as well. Although the influence of bad data quality is bigger on data-driven models.

Next, the interpretation of the model is a disadvantage for ML models. It is important to be able to explain the prediction, especially when the prediction is aberrant from the actual value. It would be advised to look into ways in which the ML models can be better understood. Also the lack of explainability is associated with a ML model. To what extend a ML model can be explained depends on the type of model that is used. For example a RF model can provide some more insights in the decisions of the model than for example a LSTM model. This problem is also present at the water authority for the current model, since this model is very complex and not always fully understood.

Another point is that currently there is not a lot of knowledge in the organization on these type of models, this is required in order to develop a ML model. The Water Authority prefers to have the knowledge on the models themselves, and not in the hands of external parties.

The current DSS operates in case of the 'daily' situation, meaning that the model should perform well during these days. In case of extreme situations, with a lot of rain or a long period without rain, the water operators take over control of the system. If the ML model is able to recognize extreme situations and predict the discharges also in an extreme situation because it has learned this during training, this would be an additional improvement in comparison to the current model. Nevertheless, the water operators think that the operational water management will never be without human-supervision, so these moments will never fully depend on only model outcomes.

4.6.3. Suggested steps

Since there is a lot of debate going on about the implementation of machine learning models and whether it is time to trust it, it will take some time before the new models will be in use. In the organization of Delfland there is the willingness to look into the possibilities, but there are some points that hinder the development of machine learning models at the water authority. The current model has been used for a long time and has proven that it works, so the urge to change is not very urgent right now.

Looking at the different barriers of the machine learning models, some recommendations are made for the organization of Delfland. First of all, the data-quality could be improved in some cases, since there is a lack of trust in the data, which is essential for a model to perform well. It would be advised to improve the quality and availability of the data and do automatic data-validation and to check the equipment by looking at the actual values and the registered values. Another advice would be to extend the knowledge on data-driven models and to find new people who can work on the development of these.

Another important step is to increase the explainability and interpretability of the data-driven models. For example by using a hybrid modelling approach (combine physics with a data-driven model) or study feature importances, further discussed in the recommendations.

Another advise to speed up the development would be to look into possibilities in which the ML model not only improves the prediction of the discharges, but also serves other purposes. There is a lot of interest from municipalities and politics in making the water system more 'future-proof'. With a machine learning model it is more feasible to take into account sustainability than in the current model. For example by taking into account the electricity market and optimizing the pump operation based on the demand and supply, which reduces the costs, but is also more sustainable. Another important goal is to improve the water quality and improve the ecology in and around the water, which can be taken into account when constructing a model as well. If these additional objectives are implemented in the ML model, the demand for these models will increase.

4.6.4. Discussion and recommendations

It must be noted that these data-driven models can only be developed if there is data available. Once this requirement is met, the development of data-driven models can be accelerated, as the demand from the government and the municipality for sustainable pumping is high. In this thesis some promising results for the prediction of the volumes of the discharges are found.

Trust and responsibility

To test the current models, it would be interesting to see what the model would predict if a test year has some outliers or unusually high values for precipitation. Next to that, some efforts need to be made to convert the prediction of the discharges in the next few hours to an input that can be used for the DSS.

The water authority is responsible for the safety of the area and need to justify the decisions that are made. Since the area is mainly located at an elevation lower than the sea level, this is crucial for the safety of the people living here, and a very big responsibility, also considering the population density and the high economic importance of the area. The water authority needs to be able to explain their decisions in all cases. Meaning that the models on which these decisions are based, need to be explainable. On top of that, the digital security of the system needs to be safe, to prevent unauthorized access, which could in the worst case lead to a big flooding disaster.

The question remains whether an organization will trust the technology that is able to learn patterns from the data without any knowledge from the system. Trust is a very important aspect for implementing a new type of model. If people do not have confidence in the performance of a model, despite good model results in a study or research, they will not implement it. A few very important aspects for improving trust in a model are the robustness of a model, the explainability and interpretability of the model. On top of that, a good data quality and availability are essential for a good performance of the model.

The trust in the model could be improved by examining moments in which the performance of the model is low by studying why at some points the model is not performing well. This will increase the trust in the model. Another method to increase the trust is by constraining the output space to only feasible outcomes. Consequently, also in case of unrealistic data, because of measurement errors or extreme situations, the model will produce realistic outputs.

Extension of the study in operational water management

The discussed steps and recommendations are found by looking at this study area and consulting the water authority of Delfland, but could be considered for other polder areas as well. However, it is strongly suggested to consult more water authorities to obtain more insights in the required steps for implementation. This will provide a broader view on the application of ML models in operational water management. On top of that, the collaboration with other water authorities could further improve the generalizability of the ML models and could help in for example obtaining additional insights in the not measured inlets.

The interviews that have been done with people from the organization have provided some valuable insights, though they are not a representative sample of the whole organization or all the water authorities in the Netherlands. The conclusions and points taken from these interviews are not always objective and have to be used with caution.

Black-box

The important consideration is the interpretability of a ML model in comparison to a physically based model. At Delfland a physical model (Sobek RR) is used to estimate the discharges from the polders. This model is not very easy or straightforward and not understood by everyone, in theory it is explainable, but this is not the case for everyone. So there should be some nuance in comparing the model, since the current physical model is also a black-box for many people. But then another question is, when is a model a black-box? Is it if you do not understand the equations in the box, or when you cannot explain why a certain prediction is wrong? Or simply when you cannot change the model if it is predicting wrong outcomes. The definition might be slightly different for everyone. Though the lack of interpretability and explainability remains a challenge for data-driven models, but there are several things that can be done to improve these disadvantages. For example by providing examples of the predictions of the model during some (extreme) events. Another option to improve the explainability and interpretability is to combine a data-driven model with a physical model, as also discussed in the subsection 4.1.3.2.

As discussed in this chapter, for implementing a new type of model, an important model property is that it can be explained why a model has made certain decisions. The more ‘deep’ a model is, the more the model itself will learn from the data, without giving back insights in on why a prediction is made or how the model exactly works. There are a few ways to overcome this ‘black box’ nature of a data-driven model. First of all by constructing more ’transparent’ models, as for example random forests. Another way to transform a black-box model to a more comprehensive model is to perform a more in depth feature importance analysis. For example by SHapley Additive exPlanations (SHAP) (Mahardhika & Putriani, 2023), to understand which features are most influential in the model’s predictions. with this method the degree of influence of each feature on the output can be calculated. Another commonly used method to increase the explainability of a ML model is LIME (Local Interpretable Model-agnostic Explanations). LIME provides a measure of feature importance for a given prediction (Dieber & Kirrane, 2020).

5 Discussion and recommendations

The possibilities of using machine learning for the prediction of the discharges has been studied, both technically and socially. In this Chapter general limitations of the research are presented and the recommendations for further research are summarized for each subquestion.

5.1. Limitations of the research

The limitations that are discussed include the physical constraints, the model performance and the characteristics of the data-driven models.

It is important to note that a data-driven model is not restricted by physical constraints, which is why it is important to have complete data to avoid the violation of the physical laws. In the present case however, information is missing about the inlets from the boezem canals to the polders. This limits the possibilities for taking into account the water balance. Another limitation is that the interactions between the polders are not taken into account, and the whole polder is estimated to be one bucket. Another point is that every polder has a different size, land use and interaction with the surrounding polders, which influences the set-up of a water balance. Next to that, time series of the data can be incomplete, which can lead to that some trends or patterns are not recognized, resulting in a lower accuracy and reliability of the model outcomes.

Next, it has to be noted that the results and conclusions are based on four pumping stations, limiting the representativeness for all the pumping stations in the area. Next to that, data-driven models are difficult to interpret, especially in the case of DL models. In subquestion 6 (Section 4.6) some recommendations are given on how to obtain more insight in the way a ML model makes predictions. Physically based models are generally easier to interpret because they are based on known physical principles, which make it easier to explain the decisions of a model. Though the lack of interpretability is a typical characteristic of data-driven models, it has to be noted that for some of the water operators, the physical and process based models can be seen as 'black-box' models as well.

Another limitation of the current models is that there is room for improvement in terms of achieving higher performance levels. This is partly caused by a limited number of trials for the optimization of the hyperparameters that have been done. In section 4.4 it is observed that for some of the hyperparameters the convergence towards an optimal value is not clear.

The societal part of this study, focusing on the practical implementation of ML models is based on a few representatives in this organization, which is not representative for all water authorities. On top of that, this research has been done at the 'Hoogheemraadschap van Delfland', which has other views on ML than other water authorities. This limits the generalizability and applicability of this research for the other water authorities and water operation systems.

5.2. Recommendations for further research

In this section a summary of the main recommendation is given for every subquestion. These recommendations have been discussed in more detail in the results chapter.

SQ1: Important considerations for modelling the discharge
If the water level is modelled, there are some notes to take into account such as the changes in the system and human influences such as irrigation and pre-pumping. If the objective is to estimate the discharge, the water levels need to be converted, which leads to some difficulties. To overcome this, it is recommended to model the discharges directly or to construct a model to make this conversion. Other recommendations are to use different modelling approaches, examples are using a hybrid approach (combine physics with a data-driven model), predict the water level deviation from the reference or to remove the pumping operation time steps in order to model only the 'natural response'.

SQ2: Estimation of the inlets
For future research it would be advised to look into the data of more polders and water authorities to get a better idea of the inlets in the polders. It could be interesting to set up an empirical formula which can be used as additional input to the model. Additionally, this empirical formula can be based on the type of polder or other characteristics if the data of also other water authorities is used. This empirical formula could be validated if for a few 'typical' polders the inlets will be measured for a given period of time.

SQ3: Feature selection and concentration time
To increase the model performance it is advised to add more features, such as the soil moisture content, nowcasting data, add higher temporal and spatial resolution evaporation data. Another recommendation would be to implement polder characteristics to increase the transferability of the models. Next it would be interesting to study the difference in radar and rain gauge data, because the difference in importance for the RF model of these variables is not fully understood. Another recommendation for the precipitation data is to use real forecasts or add noise to the shifted observations, since the current 'forecasts' are far more accurate than in reality, since they are based on the observations.

SQ4: Optimization of the hyper parameters
It is recommended to run more trials for Optuna optimization to improve the performance of the models. Next to that, performing a k-fold cross validation for the LSTM model would be recommended to see if this could further increase the performance of the model. It would be strongly suggested to compare the performance of both the optimized as the standard models in order to see the change in performance and to get an idea of the added value of the optimization. Another recommendation is to try different types of LSTM networks, such as an LSTM-CNN or a LSTM based step-sequence (LSTM-SS) framework, and to compare the performance to a 'simple' LSTM.

SQ5: Performance of the RF and LSTM model
The models are evaluated by looking at the difference between the modelled and the observed values, The difference is expressed in the MAE and the RMSE. The random forest model demonstrated very good performance, slightly surpassing the in-development machine learning model of Delfland (ReRengAI). Additionally, the random forest model outperformed the Sobek model. The Sobek model tended to overestimate discharge in the initial months. The random forest prediction of the cumulative discharge through the year accurately follows the observed cumulative discharge for the 2, 8 and 12 hour ahead prediction, but the ReRengAI model was slightly better for the 8 and 12 hourly sum. If the cumulative value at the end of the test year is considered, the ReRengAI model performed very well. The consideration that has to be made is whether this final cumulative value is more important, or all the predictions separately. The final number of the cumulative sum of the discharge is not very representative for the evaluation of the model since it does not contain any information about the performance of the model through the year. It could be possible that the model is performing very bad at the beginning by underestimating the discharge, while this is compensated in the second part of the year, resulting in an accurate final value, but a very bad actual model performance through the year.

Since the ML models fully depend on the quality of the input data, it is recommended to perform an automatic data validation and to check the distribution of the inputs of the model. Another recommendation for making a fair comparison between the constructed models and the benchmark models is to make sure the models are trained to predict exactly the same output variable and the same features are used for the predictions. Also the split in training, validation and test data should be the same and whether k-fold is used or not should be the same for a fair comparison. Additionally to the prediction of the 12 hourly discharge sum, it would be advised to predict the 2 and 8 hourly discharge sum. It would also be recommended to look into the possibilities which could further improve both models, since as in any research, there are suggestions about the used model approach.

The naïve model, which used the past 2, 8 and 12 hours for the prediction of the next, showed good results as well. This might be because the system is in many cases not changing a lot over a few days, this means the pumped discharges are predictable and could also be estimated with a simpler model. It is recommended to study the performance of simpler models and to see what the added performance is of using more complex models for modelling the discharges.

The recommendations for improving the performance of the models are discussed in detail in section 4.3.5 and 4.5.5. Additionally, it would be advised to compare the errors to the actual inputs of the DSS and see what the effect of the improved predictions is on the operation of the boezem pumps.

SQ6: Implementation of a ML model

The major recommendation regarding the implementation of a ML model is to talk to more people at the water authority and to involve more water authorities. Next, it is recommended that the data is of high quality and that the knowledge on ML models at the water authorities is extended. On top of that it is important that the explainability and interpretability of the models will be increased, for example by studying the feature importances with for example SHAP or LIME. Other options for increasing the explainability and interpretability of the model are using data-driven models that are easier to interpret or implementing physical laws in the models, which could constrain the output to feasible outcomes only. In order to accelerate the development of ML models in the organization, is would be advised to look into the ways these models can be serving several purposes.

6 Conclusion

In this research the main research question was: ”What is the performance and the added value of using machine learning for the prediction of discharges from the polders to the boezem?”. The answer is found by answering the different subquestions.

SQ1: Which factors need to be taken into account for modelling the discharge in Delfland and which time horizon is relevant? In case that the water level in the polder canals is modelled, there are some important considerations to take into account. First of all, the changes in the system like the reference water level (’peilbesluit’) and the change in water level through the year. In addition to that, the inlets from the boezem, prepumping, and irrigation, are not measured, but are important events that need to be taken into account.

It is considered to model the water level, since the pumping stations operate on the basis of a set water level threshold. However, it has been found that the conversion of water levels into discharges also posed some challenges. First, because of the different capacities at which some pumping stations operate. In addition, the pumping station does not always operate at the same water levels due to prepumping or changing the ’inslag’ and ’uitslag’ peil for example. Hence the discharges are modelled directly, so no conversion was needed. The stepwise pattern of pumped discharges posed a challenge for both models. Since the exact timing of the operation of the pumping station is not very important, it has been chosen to predict the sum of the discharges in the coming 2, 8 and 12 hours. These time horizons are chosen because knowing the discharge very far in advance cannot immediately be used due to the limited deviation from the reference water level in the system.

SQ2: How can the inlets in the system be estimated, and how are the different polders connected? In the polders water is let in regularly in order to maintain or improve the water quality in the polder canals and assure the water availability in the dry periods. This is done at many locations and these inlets are not being measured. This poses a challenge if one wants to make a data-driven model, since this crucial data is missing. Therefore the inlets in the system have been estimated by setting up a water balance for each polder. It is assumed that there have been no major changes in the storage of the system, meaning that the water balance should close because of the conservation of mass. The total amount of inlets and the seepage between the polders can be estimated by setting this change in volume in a polder to zero. The sum of the inlets and seepage differ for every polder and every year. This sum can give an idea of the order of magnitude of the inlets. The sum of these inlets and seepage was highest in the years 2018 and 2022, which can be explained by the fact that both these years were considered dry, looking at the precipitation deficit in these years. There are no connections between the polders in the considered case area, meaning that these polders do not exchange water. In other polders, this might be different, but the amount of water that is exchanged will be minimal, due to the small differences in reference water level and the lack of height differences in the area.

SQ3: Which features are important for the prediction of the discharges? For the random forest regression, the feature selection is an essential part of the model set-up. It has been found that the precipitation (radar and rain gauge data) has the greatest influence on the hourly pumped discharges up to 24 hours back in time. That is why the values for the rolling sum are added for the past 1, 2, 8 and 24 hours. Next to that, some lagged features of the water level are added to give the model an idea of the gradient of this variable. The water level observations of the previous 15 minutes, 1, 2, 4 and 8 hours are added as lagged values to provide information about the state of the system. An analysis was conducted to assess the significance of various features, considering all pumping stations and time frames. The feature importance examination of the RF model highlighted key factors crucial for discharge prediction. The most critical features identified were observed water level, discharge, predicted rainfall and the rain gauge sum of past 24 hours. Examining various time frames reveals a distinct pattern. The significance of features, apart from water level and discharge, becomes more pronounced as the prediction horizon expands. This can be explained by the longer the number of hours that will be predicted, the more important the system's current state is, and the correlation between output variable and the historical water levels and discharges diminishes. The features for the LSTM model are pre-selected using the knowledge that has been obtained with the RF model. The two most important features in the case area turned out to be the discharge and the water level, and are included for the prediction. It was decided to include the evapotranspiration feature in the LSTM model, since this is one of the components of the water balance and directly influences the amount of water in the polder. On top of that, also the precipitation has has been added as input to the LSTM model.

SQ4: What are the optimal hyper parameters of a random forest and a LSTM model? The optimal hyperparameter set differs for each pumping station and for each time horizon. Although some hyperparameters are always in the same range, no matter the pumping station or the number of hours that is predicted. If the graph of the different hyperparameter set in relation to the MAEs is studied, it can be seen that for example the parameter `max_depth` clearly converges to a certain optimum. This is one of the four hyperparameters that has been optimized and defines the maximum depth of each decision tree in the forest. This means that for future hyperparameter optimizations, the search space of this parameter could be narrowed down, which reduces the computational time. Due to the different pump capacities, the mean absolute errors for the different pumping stations vary. Despite these different capacities, in many cases the most optimal hyperparameters for the different models are not differing a lot from each other.

SQ5: What is the performance of a RF and LSTM model in comparison to the current Sobek RR model and the ReRengAI model? Random forest models have been created for the four different pumping stations to predict the sum of the discharge in the next 2, 8 and 12 hours. The LSTM model is made to predict the discharge in the next 12 hours. The performance is evaluated looking at the RMSE of the models in the test year and by considering the cumulative value of the discharges in the end of this year. Both models are compared with different benchmark models. Three benchmark models are; the Sobek RR model, the ReRengAI model and the naïve model. Considering the prediction of the discharges at node 49, the 2 and 8 hourly sum is predicted best with the RF model with a RMSE of 1,277 m^3 for the 2 hourly sum and 6,924 m^3 for the 8 hourly sum. For the prediction of the 12 hourly sum the LSTM model is slightly better with a RMSE of 10,181 m^3 in comparison to the RMSE of 11,071 m^3 of the RF. The RF model underestimates the total discharge for the prediction of the 12 hourly sum, while the cumulative sum of the LSTM model is very close to the actual observed discharges through the whole test year. During rainy days the error is generally higher for all the models. However this difference was most clear for the Sobek RR model, meaning that during the most important time steps, the model is the least accurate, which is an additional disadvantage of the current model.

SQ6: Which steps must be taken to implement a ML model for the operational control of the water in Delfland? To implement a machine learning model to predict polder discharges, there are some important notes that have to be taken into account. One requirement is that there should be enough expertise of these type of models present in the organization itself. Another important factor that needs to be considered for the implementation of a ML model is the quality of the data, there should be sufficient and high-quality data available in order to develop a reliable model. Next, the model should be robust, and also in the presence of noise and outliers the model should predict reliable outcomes.

Additionally, trusting the model and knowing what the model does, is very important to implement a ML model and will be one of the hardest challenges in the implementation. It is important that the model reacts to the observed and predicted precipitation, since this will increase the trust in the model.

"What is the performance and the added value of using machine learning for the prediction of discharges from the polders to the boezem?"

The use of machine learning models improves the performance of the prediction of the discharges from the polders to the boezem. Especially if these models are compared to the current Sobek RR model, the error decreases with more than 50% for all time horizons. This increases the accuracy of the input to the DSS system that operates the boezem pumps. Another added value of using a machine learning model is that the spatial resolution of the pumped discharges is improved, in comparison to the clustered Sobek RR model. With the more accurate estimation of the discharges, it will be easier to manage the water level in the boezem canals closer to the target water level of -0.43 m NAP. For implementing a ML model in the operational water management of a water authority the recommended steps are increasing the data-quality and improving the knowledge on ML models in the organization. Next to that, it is recommended to look into ways that the model can optimize several purposes, such as taking into account the electricity prices and the water quality for the operation of the pumps.

References

1.13. Feature selection — scikit-learn 1.2.2 documentation. (n.d.). https://scikit-learn.org/stable/modules/feature_selection.html

About Delft-FEWS - Delft-FEWS - oss.deltares.nl. (n.d.). https://oss.deltares.nl/web/delft-fews/about-delft-fews

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (n.d.). Optuna: A Next-generation Hyperparameter Optimization Framework. https://doi.org/10.1145/3292500.3330701

ArcGIS Hub. (2021, April). https://hub.arcgis.com/maps/975552a98c8241b39d531b0a0b98a78f

Bestand bodemgebruik. (2017). https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/bestand-bodemgebruik

Breiman, L. (2001). *Random Forests* (tech. rep.).

Daoud, J. I. (2017). Multicollinearity and regression analysis. *Journal of Physics: Conference Series*, *949*(1), 012009. https://doi.org/10.1088/1742-6596/949/1/012009

De Lange, R., & Koomen, A. (2023, June). Arcadis neerslagtekort app.

Delfland, H. (2022). Data hoogheemraadschap delfland [QGis]. *Hoogheemraadschap Delfland*.

Dieber, J., & Kirrane, S. (2020). Why model why? assessing the strengths and limitations of lime. *arXiv preprint arXiv:2012.00093*.

Domingos, P. (2012). review articles Tapping into the ”folk knowledge” needed to advance machine learning applications. *55*(10). https://doi.org/10.1145/2347736.2347755

Dutra, E., & Orth, R. (2020). Robustness of Process-Based versus Data-Driven Modeling in Changing Climatic Conditions. *JOURNAL OF HYDROMETEOROLOGY ams*, *21*. https://doi.org/10.1175/JHM-D

Frame, J. M., Kratzert, F., Klotz, D., Gauch, M., Shalev, G., Gilon, O., Qualls, L. M., Gupta, H. V., & Nearing, G. S. (2022). Deep learning rainfall-runoff predictions of extreme events. *Hydrology and Earth System Sciences*, *26*(13), 3377–3392. https://doi.org/10.5194/HESS-26-3377-2022

Fu, G., Jin, Y., Sun, S., Yuan, Z., & Butler, D. (2022, September). The role of deep learning in urban water management: A critical review. https://doi.org/10.1016/j.watres.2022.118973

Google. (2023, June). Keras: The high-level API for TensorFlow. https://www.tensorflow.org/guide/keras

Greenland, S., Mansournia, M. A., & Altman, D. G. (2016). Sparse data bias: A problem hiding in plain sight [Copyright - Copyright BMJ Publishing Group LTD Apr 27, 2016]. *BMJ : British Medical Journal (Online)*, *352*. https://www.proquest.com/scholarly-journals/sparse-data-bias-problem-hiding-plain-sight/docview/1785713454/se-2

Hao, R., & Bai, Z. (2023). Comparative Study for Daily Streamflow Simulation with Different Machine Learning Methods. *Water 2023, Vol. 15, Page 1179*, *15*(6), 1179. https://doi.org/10.3390/W15061179

Hochreiter, S., & Schmidhuber, J. (1997). lstm_Hochreiter. *Neural Computation*, *9*(8). https://www.bioinf.jku.at/publications/older/2604.pdf

Hoogheemraadschap Delfland. (2017). *Beleidsnota peilbeheer* (tech. rep.). https://www.hhdelfland.nl/publish/library/50/beleidsnota_peilbeheer.pdf

Hu, K. (2020). Become competent within one day in generating boxplots and violin plots for a novice without prior r experience. *Methods and Protocols*, *3*(4), 1–30. https://doi.org/10.3390/mps3040064

Imhoff, R. O., Brauer, C. C., van Heeringen, K. J., Uijlenhoet, R., & Weerts, A. H. (2022). Large-Sample Evaluation of Radar Rainfall Nowcasting for Flood Early Warning. *Water Resources Research*, *58*(3). https://doi.org/10.1029/2021WR031591

Kan, G., Yao, C., Li, Q., Li, Z., Yu, Z., Liu, Z., Ding, L., He, X., & Liang, K. (n.d.). Improving event-based rainfall-runoff simulation using an ensemble artificial neural network based hybrid data-driven model. https://doi.org/10.1007/s00477-015-1040-6

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. *Nature Reviews Physics 2021 3:6*, *3*(6), 422–440. https://doi.org/10.1038/s42254-021-00314-5

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., & Kumar, V. (2017). Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, *29*(10), 2318–2331. https://doi.org/10.1109/TKDE.2017.2720168

KNMI. (2022). Climate Explorer: Time series. https://climexp.knmi.nl/selectyear.cgi

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall-runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrology and Earth System Sciences*, *22*(11), 6005–6022. https://doi.org/10.5194/hess-22-6005-2018

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019). Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning. *Water Resources Research*, *55*(12), 11344–11354. https://doi.org/10.1029/2019WR026065

Kratzert, F., Klotz, D., Hochreiter, S., Shalev, G., Klambauer, G., & Nearing, G. (n.d.). Benchmarking a Catchment-Aware Long Short-Term Memory Network (LSTM) for Large-Scale Hydrological Modeling NeuralHydrology-Using LSTMs for rainfall-runoff modeling View project 3D Object Detection View project Benchmarking a Catchment-Aware Long Short-Term Memory Network (LSTM) for Large-Scale Hydrological Modeling. https://doi.org/10.13140/RG.2.2.18385.48487

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, *23*(12), 5089–5110. https://doi.org/10.5194/hess-23-5089-2019

Lara-Benítez, P., Carranza-García, M., & Riquelme, J. C. (2021). An Experimental Review on Deep Learning Architectures for Time Series Forecasting. *International Journal of Neural Systems*, *31*(3), 2130001. https://doi.org/10.1142/S0129065721300011

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/NATURE14539

Li, M., Zhang, Y., Wallace, J., & Campbell, E. (2020). Estimating annual runoff in response to forest change: A statistical method based on random forest. https://doi.org/10.1016/j.jhydrol.2020.125168

Li, X., Xu, W., Ren, M., Jiang, Y., & Fu, G. (2022). Hybrid CNN-LSTM models for river flow prediction. *Water Supply*, *22*(5), 4902–4919. https://doi.org/10.2166/ws.2022.170

Lim, Y. (2022, March). State-of-the-Art Machine Learning Hyperparameter Optimization with Optuna | by Yenwee Lim | Towards Data Science. https://towardsdatascience.com/state-of-the-art-machine-learning-hyperparameter-optimization-with-optuna-a315d8564de1

Lü, H., Hou, T., Horton, R., Zhu, Y., Chen, X., Jia, Y., Wang, W., & Fu, X. (2012). The streamflow estimation using the Xinanjiang rainfall runoff model and dual state-parameter estimation method. https://doi.org/10.1016/j.jhydrol.2012.12.011

Mahardhika, S. P., & Putriani, O. (2023). Deployment and use of Artificial Intelligence (AI) in water resources and water management. *IOP Conference Series: Earth and Environmental Science*, *1195*(1). https://doi.org/10.1088/1755-1315/1195/1/012056

Muñoz, P., Orellana-Alvear, J., Willems, P., & Célleri, R. (n.d.). Flash-Flood Forecasting in an Andean Mountain Catchment-Development of a Step-Wise Methodology Based on the Random Forest Algorithm. https://doi.org/10.3390/w10111519

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Nguyen, H.-P., Liu, J., & Zio, E. (2020). A long-term prediction approach based on long short-term memory neural networks with automatic parameter optimization by Tree-structured Parzen Estimator and applied to time-series data of NPP steam generators. *Applied Soft Computing Journal*, *89*, 106116. https://doi.org/10.1016/j.asoc.2020.106116

Parisouj, P., Mokari, E., Mohebzadeh, H., Goharnejad, H., Jun, C., Oh, J., & Bateni, S. M. (2022). Physics-Informed Data-Driven Model for Predicting Streamflow: A Case Study of the Voshmgir Basin, Iran. *Applied Sciences (Switzerland)*, *12*(15). https://doi.org/10.3390/app12157464

PDOK. (2021, March). Nationaal georegister. https://www.nationaalgeoregister.nl/geonetwork/srv/dut/catalog.search#/metadata/9ad3f0c0-9e2c-4d44-a467-b57920aa512f?tab=general

Probst, P., Wright, M. N., & Boulesteix, A. L. (2019, May). Hyperparameters and tuning strategies for random forest. https://doi.org/10.1002/widm.1301

Prudden, R., Adams, S., Kangin, D., Robinson, N., Ravuri, S., Mohamed, S., & Arribas, A. (2020). A REVIEW OF RADAR-BASED NOWCASTING OF PRECIPITATION AND APPLICABLE MACHINE LEARNING TECHNIQUES A PREPRINT.

Qiao, X., Peng, T., Sun, N., Zhang, C., Liu, Q., Zhang, Y., Wang, Y., & Shahzad Nazir, M. (2023). Metaheuristic evolutionary deep learning model based on temporal convolutional network, improved aquila optimizer and random forest for rainfall-runoff simulation and multi-step runoff prediction. *Expert Systems With Applications*, *229*, 120616. https://doi.org/10.1016/j.eswa.2023.120616

Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, *378*, 686–707. https://doi.org/10.1016/j.jcp.2018.10.045

Razavi, S. (2021). Deep learning, explained: Fundamentals, explainability, and bridgeability to process-based modelling. *Environmental Modelling and Software*, *144*. https://doi.org/10.1016/j.envsoft.2021.105159

Rumelhart, D. E. (1986). Learning representations by back-propagating errors. *Institute for Cognitive Science, C-015, University of California, San Diego, La Jolla, California, 92093, USA*.

Setianto, A., & Triandini, T. (2013, June). *COMPARISON OF KRIGING AND INVERSE DISTANCE WEIGHTED (IDW) INTERPOLATION METHODS IN LINEAMENT EXTRACTION AND ANALYSIS* (tech. rep. No. 1). Geological Engineering Departement Universitas Gadjah Mada.

Shen, C. (2018, November). A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists. https://doi.org/10.1029/2018WR022643

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., Woo, W.-C., & Kong Observatory, H. (n.d.). *Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting* (tech. rep.).

Song, T., Ding, W., Liu, H., Wu, J., Zhou, H., & Chu, J. (2020). Uncertainty Quantification in Machine Learning Modeling for Multi-Step Time Series Forecasting: Example of Recurrent Neural Networks in Discharge Simulations. *Water 2020, Vol. 12, Page 912*, *12*(3), 912. https://doi.org/10.3390/W12030912

Stalknecht, A., von Meijenfeldt, B., & Gnodde, S. (2021). Eindrapportage rerengai project. *Hoogheemraadschap Delfland*.

Swischuk, R., Mainini, L., Peherstorfer, B., & Willcox, K. (2019). Projection-based model reduction: Formulations for physics-based machine learning. *Computers and Fluids*, *179*, 704–717. https://doi.org/10.1016/j.compfluid.2018.07.021

Valenzuela, O., Rojas, F., Luis, ·., Herrera, J., Pomares, H., & Rojas, I. (n.d.). *Contributions to Statistics Theory and Applications of Time Series Analysis Selected Contributions from ITISE 2019* (tech. rep.). http://www.springer.com/series/2912

Vegter, G., Vreugdenhil, H., & Jouwersma, S. (2017). Rainlevelr.

Von Rueden, L., Mayer, S., Beckh, K., Georgiev, B., Giesselbach, S., Heese, R., Kirsch, B., Pfrommer, J., Pick, A., Ramamurthy, R., Walczak, M., Garcke, J., Bauckhage, C., & Schuecker, J. (2023). Informed Machine Learning - A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems. *IEEE Transactions on Knowledge and Data Engineering*, *35*(1), 614–633. https://doi.org/10.1109/TKDE.2021.3079836

Wilbrand, K., Taormina, R., ten Veldhuis, M.-C., Visser, M., Hrachowitz, M., Nuttall, J., & Dahm, R. (2023). Predicting streamflow with lstm networks using global datasets. *Frontiers in Water*, *5*, 1166124.

Xiang, Z., & Demir, I. (n.d.). *Fully distributed rainfall-runoff modeling using spatial-temporal graph neural network* (tech. rep.).

Yin, H., Wang, F., Zhang, X., Zhang, Y., Chen, J., Xia, R., & Jin, J. (2022). Rainfall-runoff modeling using long short-term memory based step-sequence framework. *Journal of Hydrology*, *610*, 127901. https://doi.org/https://doi.org/10.1016/j.jhydrol.2022.127901

Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A review of recurrent neural networks: Lstm cells and network architectures. *Neural Computation*, *31*(7), 1235–1270. https://doi.org/10.1162/NECO{_}A{_}01199

Yuswo, M. (2022). Drainage System of Tegalsari Polder for Handling Flood and Tide in Tegal City Indonesia. *IOP Conference Series: Earth and Environmental Science*. https://doi.org/10.1088/1755-1315/955/1/012008

A
Appendix

A.1. Study-area

A.1.1. Rainfall-runoff response in a polder

In a polder with (almost) no gravitational flow, there are other factors determining the speed in which the water ends up in the canals. The biggest influence on the rainfall-runoff response is thus not the slope or the distance from the stream, but the type of surface and the connection to the open water. For example, a small polder with a lot of paved surface and no infiltration, the response to a big precipitation event will be a lot faster than in a big catchment in which the water is able to infiltrate into the ground.

(Urban) Runoff to surface water

An overview of the amount of water ending up in the surface waters depends on the type of landuse and, next to that, on whether the surface is connected to a combined sewer system or that the water ends up directly in the canals. A map of the proportion of the water is obtained from Delfland. This data could be used to make an estimation of the total amount of water that goes directly into the polder- or boezemcanals, and which part goes to a waste water treatment plant.

Rainlevelr

Recently a new method for storing water in the polder is the Rainlevelr project (Vegter et al., 2017). The project is about working together with farmers with large greenhouses that have a water storage, such as a basin, where precipitation is stored. This storage is filled by the water that falls directly into the basin and the water that falls on the greenhouse. In the case of a big precipitation event, the water board reports to the company whether they could empty their storage before the rain starts. In this case the water managers can start with pumping the water out of the polder earlier. The influence on the rainfall-runoff response in recent years is minor, but it could be interesting for the future, since the scale of the project is increasing. Currently the effect is probably not very large because the current surface that is ‘covered’ by this project is 334 hectares, which means less than 1% of the total area of Delfland. Most of these water basins are located in the polders in the ‘boezemland’. These are the ‘higher’ elevated polders in the Western part of the area, in this research the focus is on the lower lying polders.

A.1.2. Boezem pumping stations

Table A.1: Boezem pumps in Delfland

'Boezem' pumps	**Remark**	**Capacity [m^3/min]**
Parksluizen	Often used for the inlet of water.	300-1000+
Zaaijer	Often used	200-1000+
Doorvoergemaal EON	Mostly pumping	150
Schiegemaal	Usually used for shorter periods	225
Dolk	supply pump from Rijnland, used during dry periods	
Schoute	Often used, sometimes long periods, but also short	300
Vlaardinger Driesluizen	Regulary pumping to driesluizen	8.25
Wateringsesluis	Not used often	10
Westland		100-1000
Winsemius (Brielse meer)	Often used in summer	180-240
V/d Burg	Often used	90

A.1.3. Polder pumping stations

In the Holierhoekse- en Zouteveensepolder and the second pumping station of the Duifpolder the pumping station operates with more than three different capacities (see Figure A.1).

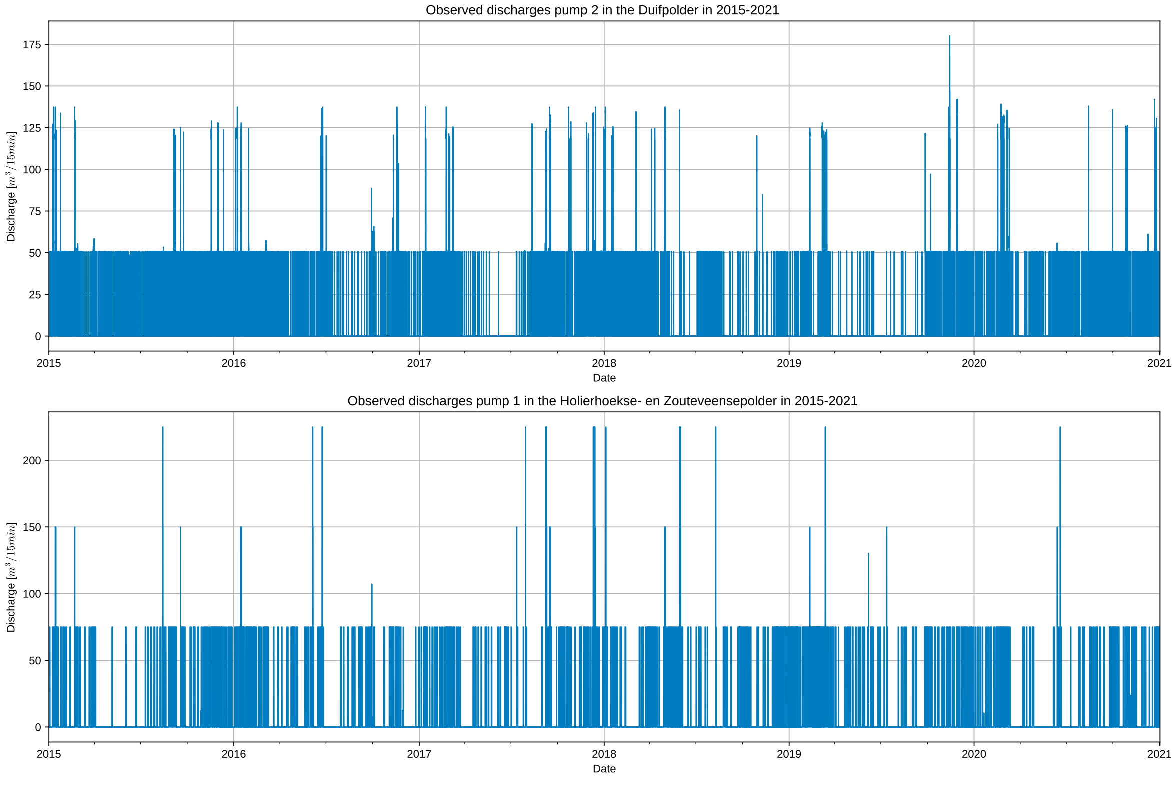


Figure A.1: The observed discharge in the period 2014-2021 at pumping station 2 in the Duifpolder and pumping station 1 in the Holierhoekse- en Zouteveense polder

A.1.4. Landuse

Delfland is an interesting area; a simplified overview of the area with different types of land use is shown in Figure A.2. Most of the area can be considered as urban / industrial area, greenhouses and rural / green area. The area of Delfland is mostly below sealevel, densely populated and of great importance for the Dutch economy.

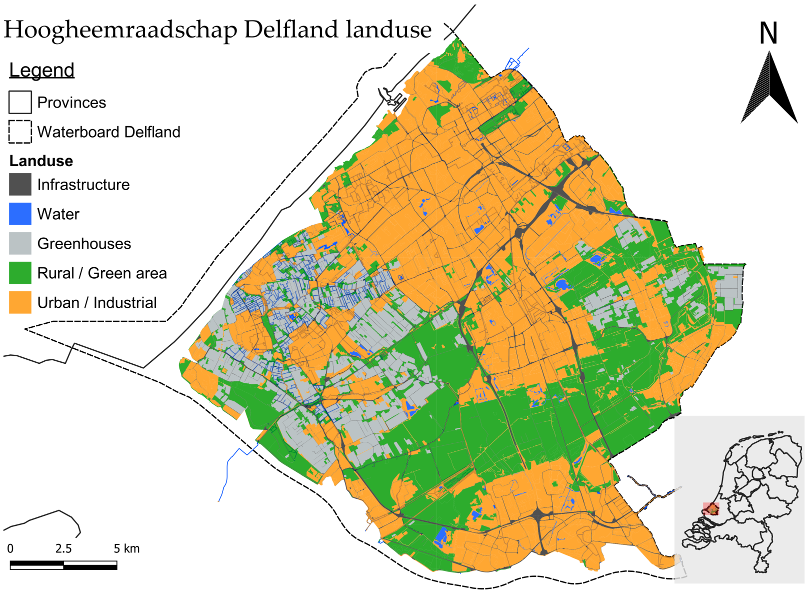


Figure A.2: Hoogheemraadschap Delfland landuse for the area on the inside of the dikes (“ArcGIS Hub”, 2021) (PDOK, 2021) (“Bestand bodemgebruik”, 2017) (Delfland, 2022)

A.2. Machine learning

In this Section more information on different terminologies around Machine Learning and neural networks are explained. Next to that, the in-development machine learning model of Delfland is discussed.

A.2.1. Artificial intelligence and machine learning

Artificial intelligence (AI) is the overarching term for all machines that use data to mimic or even beat human intelligence. Examples of the development of computer systems that can perform tasks that typically require human intelligence are speech recognition, decision-making, and language understanding. The two main types of AI are narrow (weak) or general (strong) AI. Narrow AI is designed to perform a specific task. It is highly specialized and can be very good at the task it is designed for, but it lacks the general intelligence that humans have. Examples include voice assistants like Siri or recommendation systems like those used by Netflix or Spotify. General AI is a theoretical form of AI that has the ability to understand, learn, and apply intelligence across a wide range of tasks, similar to human intelligence. General AI does not currently exist and is a subject of ongoing research.

Machine learning is an essential tool used to achieve AI. In many AI applications, machine learning algorithms are used to train models on large amounts of data to recognize patterns and make predictions or decisions. The performance increases once more data is processed. There are several types of tasks, for example, regression, classification, and clustering.

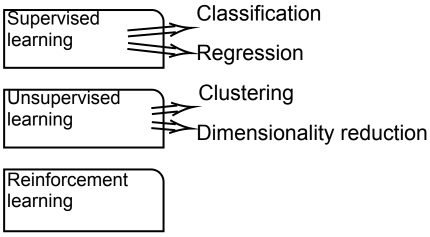


Figure A.3: Types of machine learning

The three forms of ML are supervised, unsupervised and reinforcement learning (see Figure A.3). With as main difference the way it learns from data. In supervised learning, the input data is connected to a label, for example, a dataset with images and a label corresponding to that image (house, dog or human). The label (observed target variable) can be either categorical or continuous. The machine learns from the training data and uses this knowledge to predict unlabeled data (Lecun et al., 2015). In unsupervised learning, there are no labels, this means the machine has to find patterns in the data without knowing anything about these patterns. An example could be to cluster groups of dogs based on a recognized pattern. Patterns of the given dataset, organize the data according to found patterns and similarities, and then present that information in a concise, simplified version. The last form of ML is reinforcement learning, this means the machine learns by trial and error. The machine receives feedback for every good or bad choice it makes and adjusts its decisions in order to maximize the performance.

All the machine learning models algorithms learn the following three standard steps. The first part is the 'representation', here the classifier should be expressed in a understandable 'language' for the machine. The choice of representation for a learner is equivalent to selecting the range of classifiers it can learn, known as the hypothesis space. Any classifier outside of the hypothesis space is not learnable. This is also part of the feature selection step.

The next learning subset is how to evaluate the classifier. This can be done with several evaluation metrics. Standard evaluation metrics, misclassification score (precision and recall) or the loss that is being optimized.

Then there is a optimization step. In this step a method for searching the classifiers that that scores best, which means the loss-function is minimized (Domingos, 2012).

The boost in the use of deep learning for several purposes was accelerated by the introduction of fast graphics processing units (GPUs), which increased the speed of training 10 to 20 times and were favorable to program (Lecun et al., 2015).

A.2.2. Neural networks

A neural network consists of layers of artificial neurons. Similarly as in a biological synapse, the interneuron connections are assigned weights that dictate the strength of the signal. Each neuron is associated with a bias value that simulates the activation threshold.

A type of neuron is a perceptron, which can be schematized as a small device that transfers a number of inputs into an output. The different inputs can have different weights representing the importance of that input. The output of the perceptron is determined based on whether the weighted sum is smaller than or greater than a certain threshold value, which is represented in algebraic terms below.

$$\text{output } = \begin{cases} 0 & \text{if } \sum_{\text{j}} w_{\text{j}} x_{\text{j}} \leq \text{ threshold} \\ 1 & \text{if } \sum_{\text{j}} w_{\text{j}} x_{\text{j}} > \text{ threshold} \end{cases}$$

In Figure A.4 the different layers of perceptrons are shown. Every perceptron makes a certain decision based on the inputs and generates one output. The more perceptrons, the more complex the whole process of making a decision is. In the decision-making process of a perceptron, a bias can be introduced which represents the threshold discussed above.

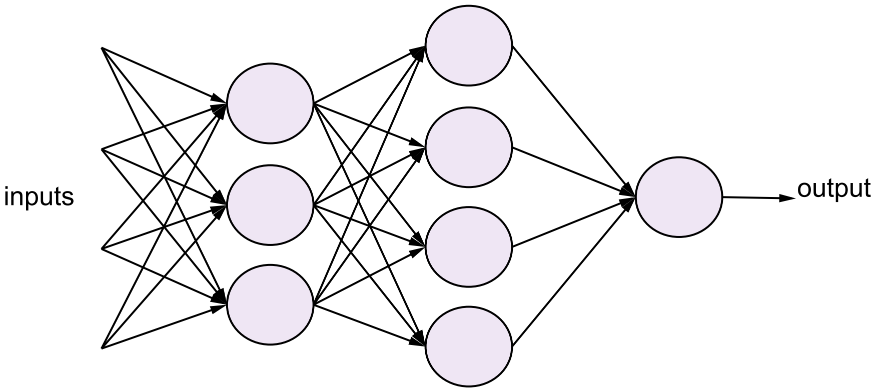


Figure A.4: A Multi-layer feed forward neural network.

The bias is a measure of the 'effort' it takes to get a perceptron to give 1 as an output.

$$\text{output } = \begin{cases} 0 & \text{if } w * x + b \leq \text{ 0} \\ 1 & \text{if } w * x + b > \text{ 0} \end{cases}$$

Equation A.1 illustrates how input information is processed in a single neuron, which is defined by the equation:

$$y = \phi \left(b + \sum_{\text{i}=1}^{\text{m}} (\omega_{\text{i}} x_{\text{i}}) \right) \tag{A.1}$$

In this equation there are *m* input variables, which are the number of features. These features are multiplied by a weight *w*.

In a perceptron, the output of a function is 0 or 1. This makes it more difficult to see how the weights and biases gradually adjust towards the desired outcome. There are different types, one is the sigmoid function.

$$\sigma(z) \equiv \frac{1}{1 + e^{-\text{z}}}$$

The output is the following:

$$\frac{1}{1 + \text{exp}\left(- \sum_{\text{j}} w_{\text{j}} x_{\text{j}} - b \right)}$$

A single neuron or synapse cannot learn or process complex information. However, a network of many neurons can learn complex tasks such as regression. Neural networks are composed of layers of neurons, including an input layer, an output layer, and at least one hidden layer with an arbitrary number of neurons. Deep neural networks have two or more hidden layers and can represent more complex patterns between input and output as the level of abstraction increases with the number of neurons and layers. However, the optimal architecture for a specific application depends on the available training data, and a deeper network or higher number of neurons per layer does not necessarily improve pattern recognition.

A set of units compute a weighted sum to go from one layer to another and this result is passed through an activition function (ϕ) such as as sigmoid, rectified linear (ReLU) or a tanh(z) function, which are non-linear functions.

Types of Neural Networks There are several types of neural networks, the main types are summarized in the table below.

- Perceptron: simple model of a neuron. Binary classifier, and processes data into two categories.
- Multilayer perceptron neural network: Fully connected network in which each node is connected to the nodes in the following layer.
- Feed-forward Neural Networks (FNN): A simple artificial Neural Network in which the data moves only in one direction without back-propagation or feedback loops. In this structure all layers are fully connected. The input is mostly tabular or text data.
- Convolutional Neural Networks (CNN): This is a combination of multilayer perceptrons and one or more convolutional layers that are fully connected or pooled. This type of NN is mainly used for image recognition and are capable of detecting features without human supervision.
- Recurrent Neural Networks (RNN): The layers in this network are connected, which means that relevant information is send back and used again as input. This improves the understanding of the context of an input and improves the prediction. These type of networks are useful for processing sequential data.

Artificial neural networks (ANNs) are simplified versions of the neurons found in the brain. Each node in an ANN is connected to other nodes, with varying densities and amounts of connections depending on the type of network. The nodes are grouped into layers between the input and output layers, forming a multi-layered network architecture known as a deep neural network. These layers enable the network to learn different features of data, with hidden layers allowing the understanding of complex patterns and concepts.

A.2.3. Machine learning developments Hoogheemraadschap Delfland

The ReRengAI project was done to look into the possibilities of using ML for predicting the pumped discharge from the polders into the boezem. In this model Light Gradient Boosting (LGB) is used. The advantage of using a LGB model is that it has high predictive power, especially in the first few hours (Stalknecht et al., 2021). The drawback of using this model is that an unique model is created for every location, meaning 432 models, which means high computational time. That is why in the current model it is decided to use this only for the first 6 hours. The last 30 hours are predicted using a linear regression, more specifically, a Ridge regression. This type of regression prevents overfitting by using L2 regularization. This model is able to compute very fast and the performance is comparable to the more complex ML model. In Figure A.5 the predicted and actual discharge of the 'Gemaal van de Eshofpolder' is shown.

The ReRengAI project is not being used to operate yet, but it is run real time to see how the ML model performs. The current DSS (BOS) is very suitable for the implementation of machine learning models.

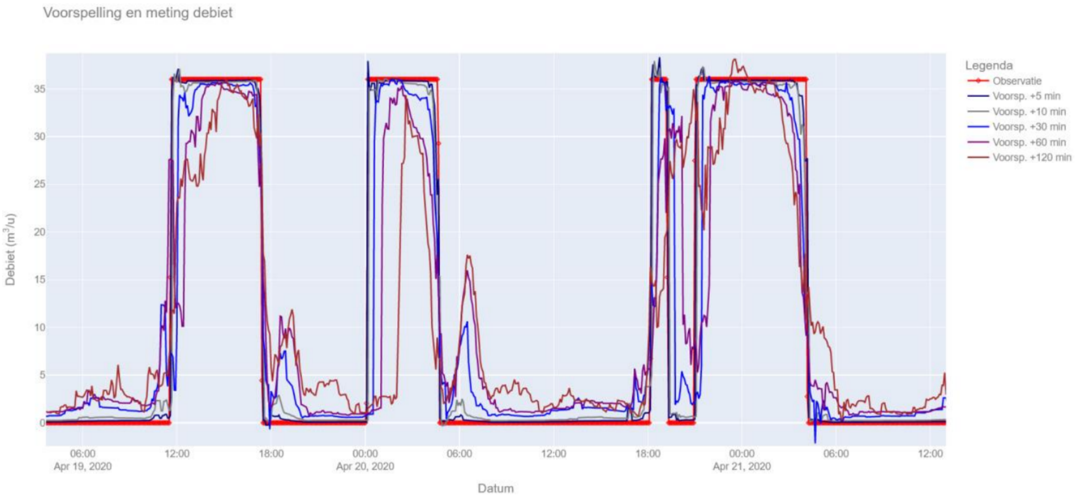


Figure A.5: Observed and predicted discharge with the ReRengAI LGB (Stalknecht et al., 2021)

The inputs of the model are the radar data, the pumped discharges, the waterlevels and the data of the rain gauges in Delfland. The pumped discharges are the output variable. For each input, the historical data of the eight previous hours is used, except for the radar data, in this case five hours historical and three hours predicted precipitation data is used. In all cases, the rolling mean is used as a value for the past half an hour. This means, with four variables, the model gets 4*16=64 features.

A.3. Available data

A.3.1. Data 'Hoogheemraadschap van Delfland'

Most of the data is provided by Sjoerd Gnodde, who works as a data scientist at the waterboard. The available data is presented in Table A.2.

The rain gauges are coupled to a certain drainage area. In total there are 10 rain gauges in the area of Delfland:

- RGN301104
- RGN306170
- RGN202146
- RGN111102
- RGN401101
- RGN210101
- RGN112103
- RGN450101
- RGN115110
- RGN101102

Table A.2: Data Hoogheemraadschap

Parameter	Filename	Interval	Unit	Start	End
Discharges poldergemalen	tijdsreeksen_gemalen.feather	5 minutes	m^3/min – average value of last 5 minutes	12-09-2012	10-10-2022
Waterlevels	tijdsreeksen_waterstanden.feather	5 minutes	meter NAP	12-09-2012	10-10-2022
Precipitation	tijdsreeksen_neerslag.feather	15 minutes	mm (cumulative mm of 15 minutes)	12-09-2012	10-10-2022
Precipitation radar (via Hydronet)	tijdsreeksen_neerslag_afvoergebied.feather	5 minutes	mm (cumulative mm of last 5 minutes)	12-09-2012	10-10-2022
	Temp_radar_neerslag	5 minutes	mm (cummulative mm of last 5 minutes)	01-01-2012	20-04-2022

Table A.3: Data files Hoogheemraadschap

Parameter	Filename	GIS-bestand
Discharges poldergemalen	tijdsreeksen_gemalen.feather	gemalen.shp
Waterlevels	tijdsreeksen_waterstanden.feather	meetlocatie.shp
Precipitation	tijdsreeksen_neerslag.feather	neerslagmeters.shp
Precipitation radar (via Hydronet)	tijdsreeksen_neerslag_afvoergebied.feather	AfvoergebiedAanvoergebied.shp

Information on the soil moisture content is obtained by the 'Waterschapshuis' and is based on satelite measurement. This could be obtained by an API (Satdata).

A.3.2. Nowcasting data

It would be interesting to obtain nowcasting rainfall data, since this method is able to make predictions of precipitation on a timescale of less than 2 hours (Prudden et al., 2020).

A.3.3. KNMI data

Figures of the evaporation and the temperature (KNMI, 2022).

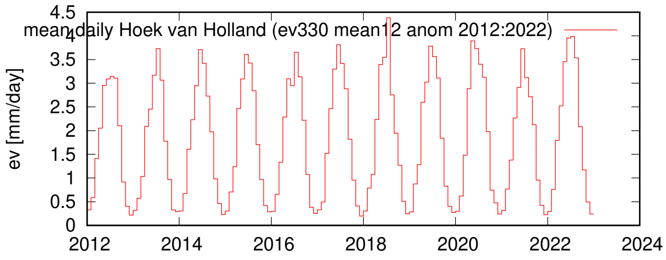


Figure A.6: Evaporation Hoek van Holland

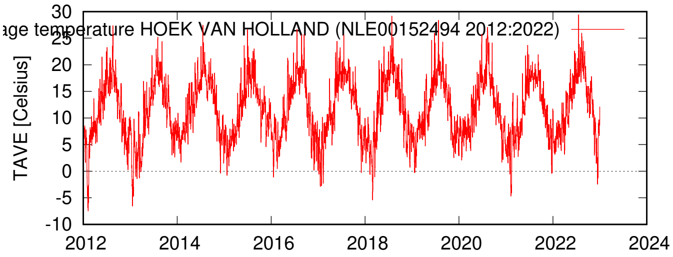


Figure A.7: Temperature Hoek van Holland

There are multiple KNMI stations located in and around Delfland, including Hoek van Holland, Rotterdam, and Voorschoten. To analyze the spatial differences between these locations, correlation plots are created for the Hoek van Holland and Rotterdam stations. Notably, a significant contrast is observed in the daily precipitation data at the two different locations, which is expected due to the highly heterogeneous nature of rain. For the prediction of water levels, the precipitation data from Delfland is utilized, and not the data from the KNMI. This data is specific for each polder and likely more accurate.

As for the other parameters, there is relatively little variation between the two different stations. Consequently, the timeseries of one station (Hoek van Holland) is considered sufficient for the whole area due to the lack of substantial differences.

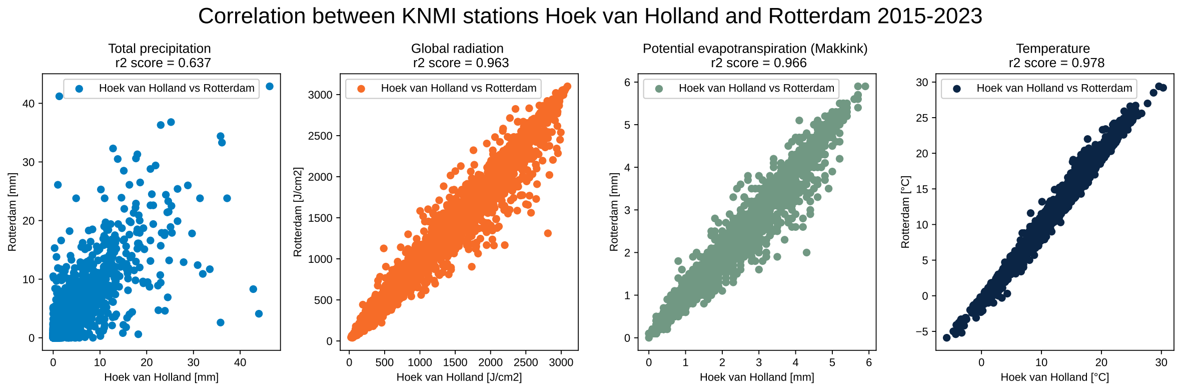


Figure A.8: Correlations between the daily observations of the KNMI stations Hoek van Holland and Rotterdam of precipitation, global radiation, temperature and Makkink evaporation.

A.3.4. Catchment specific features

Additional catchment specific features that can be added are presented in Table A.4.

Table A.4: Characteristics of the pump and catchment

	Feature	Unit	Source
Pump characteristics	Capacity	m^3/min	Hoogheemraadschap Delfland
Catchment characteristics	Area	m^2	Hoogheemraadschap Delfland
	Water level (polderpeil)	m N.A.P.	Hoogheemraadschap Delfland
	Number of pumps	[-]	Hoogheemraadschap Delfland
	Total length of canals	m	Hoogheemraadschap Delfland
	Percentage connected to sewer	%	Hoogheemraadschap Delfland
	Permeability surface	%	

A.4. Model set-up

In this section more details about the set-up of the two different models is discussed.

A.4.1. Random Forest Regression

Steps Random Forest model

The steps for the set-up of the Random Forest Regression model:

1. Load and prepare data
2. Define the features (X) (Section 3.2.2.2)
3. Target variable (y): the sum of the discharge in the next 2, 8 or 12 hours (Figure 3.5)
4. Split data into a training and test dataset.
5. Create the model and tune the hyperparameters using k-fold cross validation (Section 3.3.2.1)
6. Make predictions
7. Evaluate model and compare with benchmark models

Hyperparameters Random Forest

- `n_estimators`: This hyperparameter determines the number of decision trees to be included in the random forest. Increasing the number of estimators typically improves the model's performance, but it also increases the computational cost. It is important to find a balance between accuracy and efficiency.
- `max_depth`: This hyperparameter sets the maximum depth of each decision tree in the random forest. A larger value allows the trees to grow deeper, potentially capturing more complex patterns in the data. However, increasing max depth can also lead to overfitting, especially if the dataset is small. It is important to tune this parameter to find an appropriate balance.
- `min_samples_ split` (node size): This hyperparameter specifies the minimum number of data points required to split an internal node during tree construction. A higher value can prevent overfitting by ensuring that a node must have a sufficient number of samples before it can be split. However, setting this value too high can result in underfitting, where the trees fail to capture important patterns in the data.
- `min_samples_leaf`: This parameter sets the minimum number of samples (data points) that a leaf node must have. If, after a split, a node has fewer samples than the value set by `min_samples_leaf`, the algorithm will stop further splitting that node and make it a leaf.

A.4.2. LSTM

Steps LSTM model

The steps for the set-up of the LSTM model:

1. Load and prepare data
2. Choose the input features 3.3.2.2.
3. Target variable (y): the sum of the discharge in the next 12 hours (Figure 3.5).
4. Split data into a training, validation and test dataset.
5. Scale the data (Section 3.2.2.5).
6. Create sequences (Figure 3.9).
7. Define callbacks
8. Optimize the hyperparameters with Optuna by minimizing the objective function (Section 3.3.3).
9. Use the best hyperparameterset to train the model (Subsection 3.3.2.1).
10. Make predictions.
11. Inverse transform the input and output features with the scaler to interpret the results (Subsection 3.2.2.5).
12. Evaluate model and compare with benchmark model (Subsection 3.4).

Hyperparameters LSTM

- Number of nodes (50-200): hidden neurons
- Activation function (tanh, relu)
- Dropout rate (0 - 0.5)
- Learning rate (0.00001 - 0.1)
- Batch size (16-64): The number of samples in each batch during training and testing
- Number of hidden layers (1 - 2)
- Number of units in a dense layer
- Weight initialization
- Decay rate
- Momentum (0.5 - 0.9)
- Number of epochs
- Time steps: Input length

A.4.3. Optuna optimization

An example of the objective function that is used for the hyperparameter search for the RF model is presented in the code below.

Listing A.1: Objective function for hyperparameter optimization

```
import numpy as np
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score

def objective(trial, X_train, y_train, X_test, y_test):
    # Suggest values for hyperparameters
    n_estimators = trial.suggest_int("n_estimators", 10, 200, log=True)
    max_depth = trial.suggest_int("max_depth", 2, 32)
    min_samples_split = trial.suggest_int("min_samples_split", 2, 12)
    min_samples_leaf = trial.suggest_int("min_samples_leaf", 1, 12)

    # Create and fit random forest model
    model = RandomForestRegressor(
        n_estimators=n_estimators,
        max_depth=max_depth,
        min_samples_split=min_samples_split,
        min_samples_leaf=min_samples_leaf,
        random_state=42, n_jobs=-1)

    model.fit(X_train, y_train)

    # Make predictions and calculate performance metrics
    y_pred = model.predict(X_test)
    rmse = np.sqrt(mean_squared_error(y_test, y_pred))
    mae = mean_absolute_error(y_test, y_pred)
    r2 = r2_score(y_test, y_pred)
    return mae
```

A.5. Additional results

In this Appendix, some additional figures for the different time horizons are presented.

A.5.1. 2 hourly sum

The 2 hourly sum of the discharge per pump is presented in Figures A.9, A.10, A.11 and A.12.

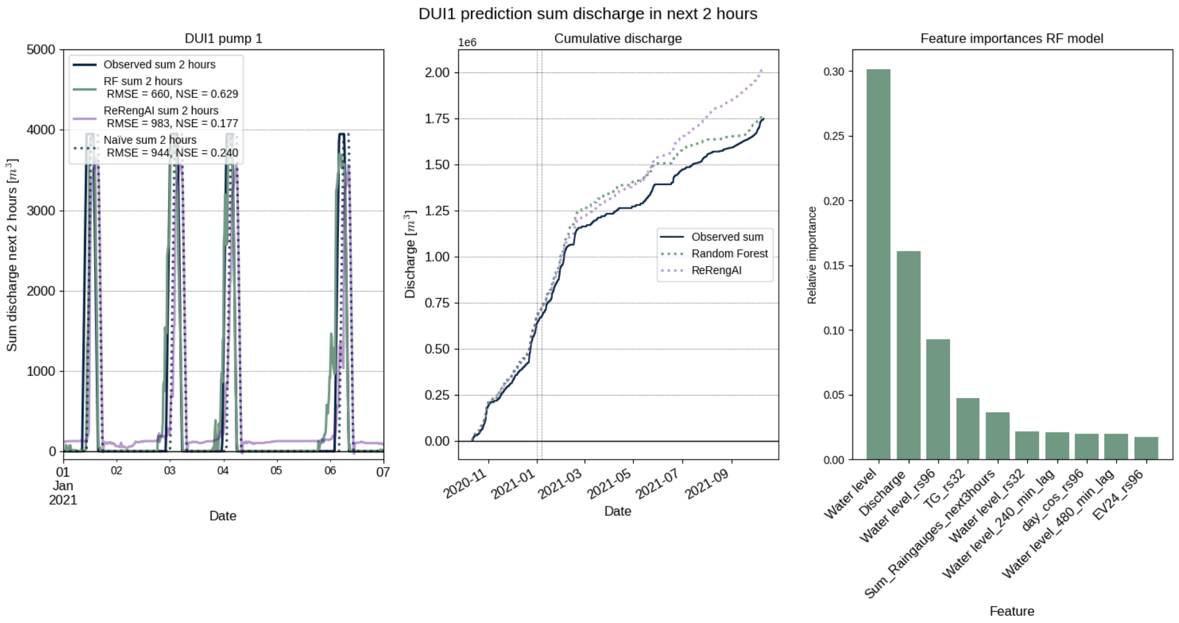


Figure A.9: Comparison Random Forest, ReRengAI, Naïve and LSTM model for pump 1 of the Duifpolder (DUI1)

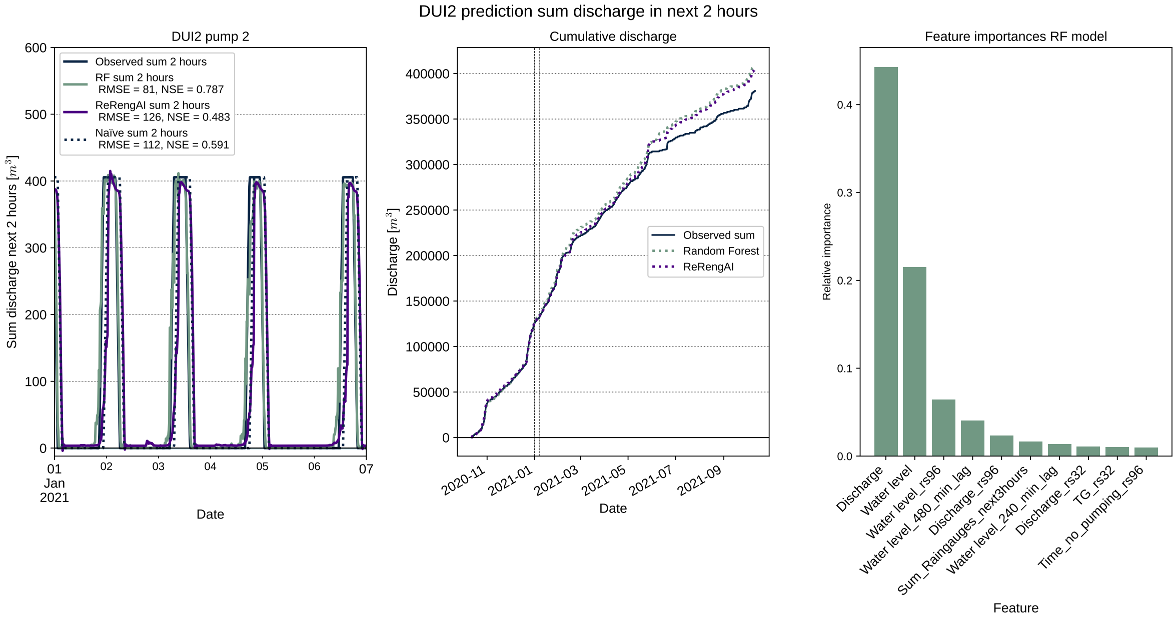


Figure A.10: Comparison Random Forest, ReRengAI, Naïve, and LSTM model for pump 2 of the Duifpolder (DUI2)

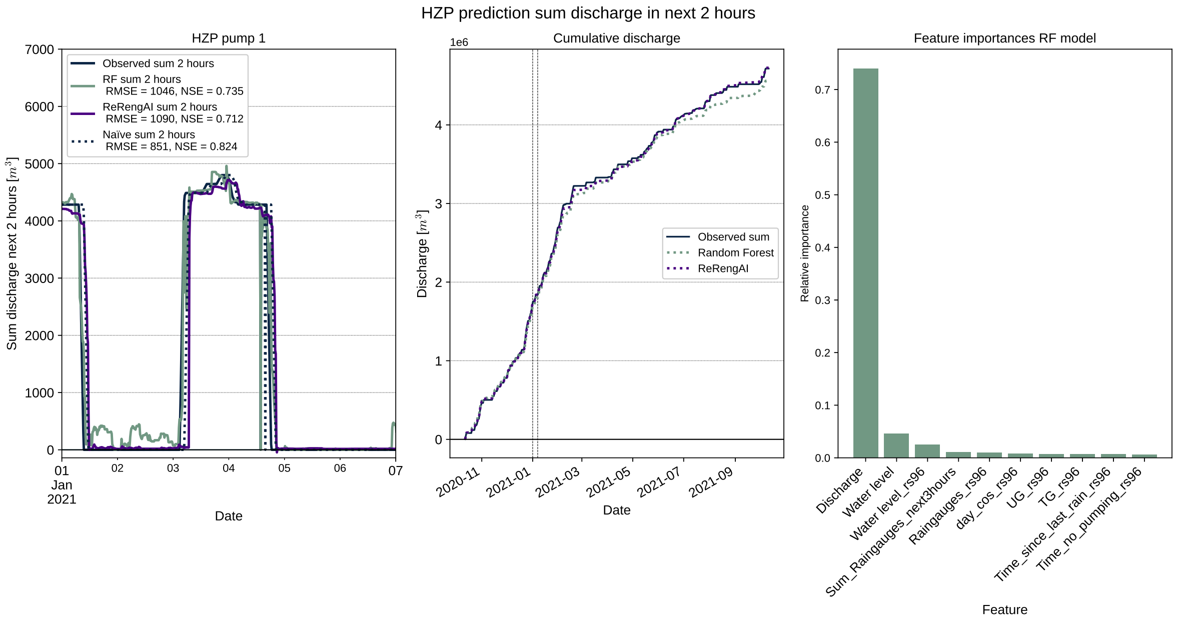


Figure A.11: Comparison Random Forest, ReRengAI, Naïve and LSTM model for the HZP

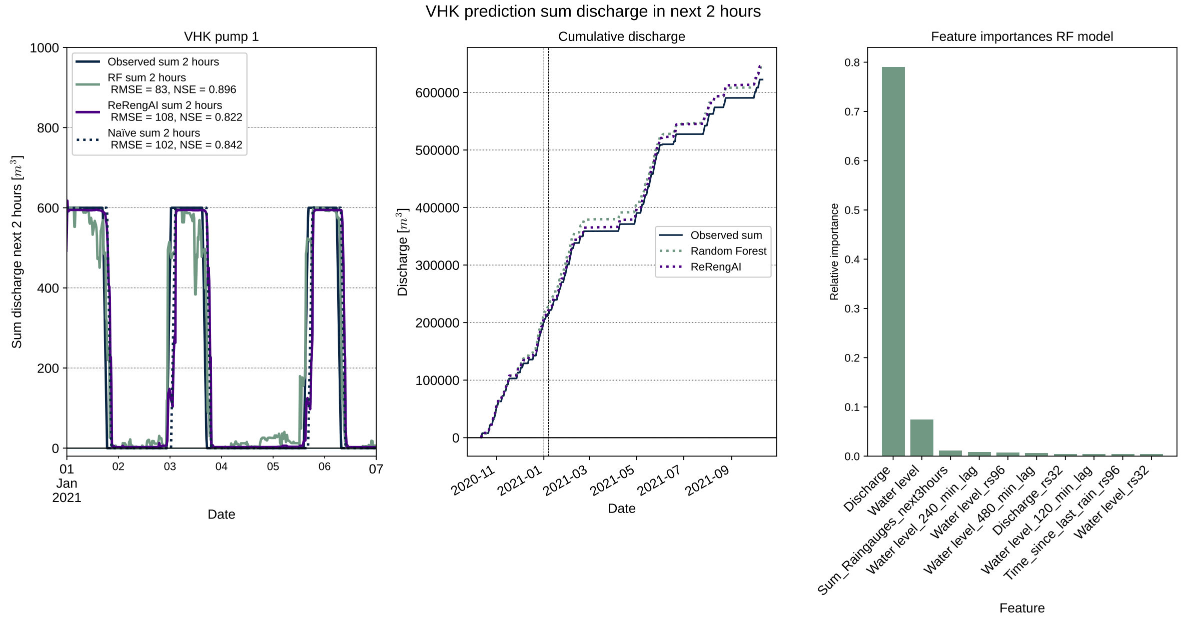


Figure A.12: Comparison Random Forest, ReRengAI, Naïve, and LSTM model for the VHK

A.5.2. 12 hourly sum

The 12 hourly sum of the discharge per pump is presented in Figure A.13, A.14, A.15, A.16.

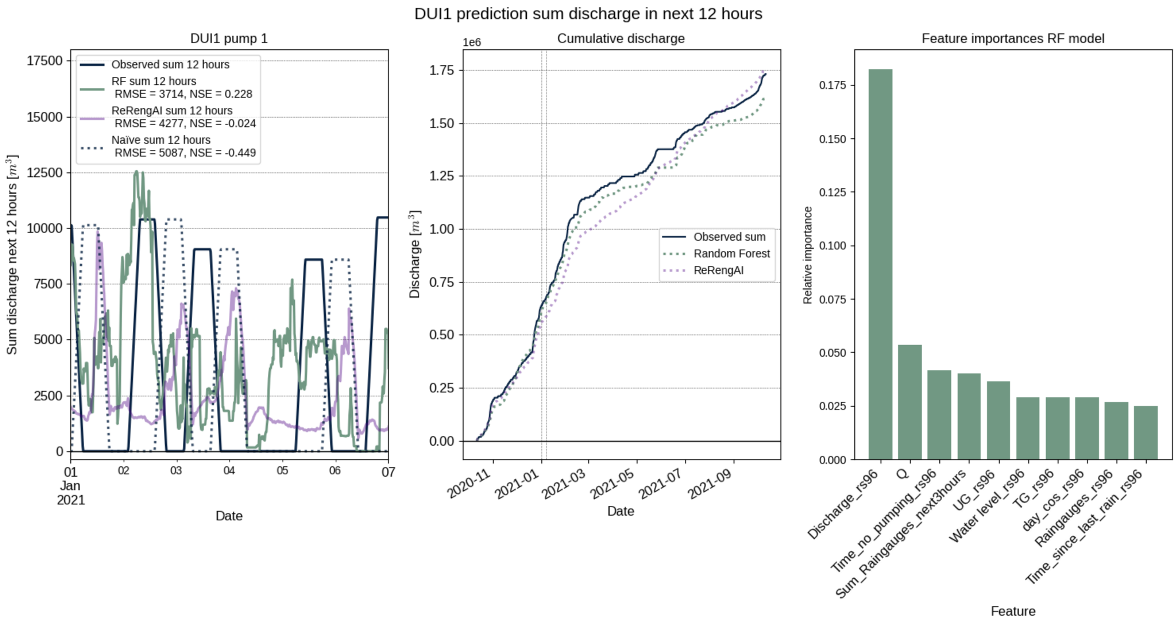


Figure A.13: Comparison Random Forest, ReRengAI, Naïve and LSTM model for pump 1 of the Duifpolder (DUI1)

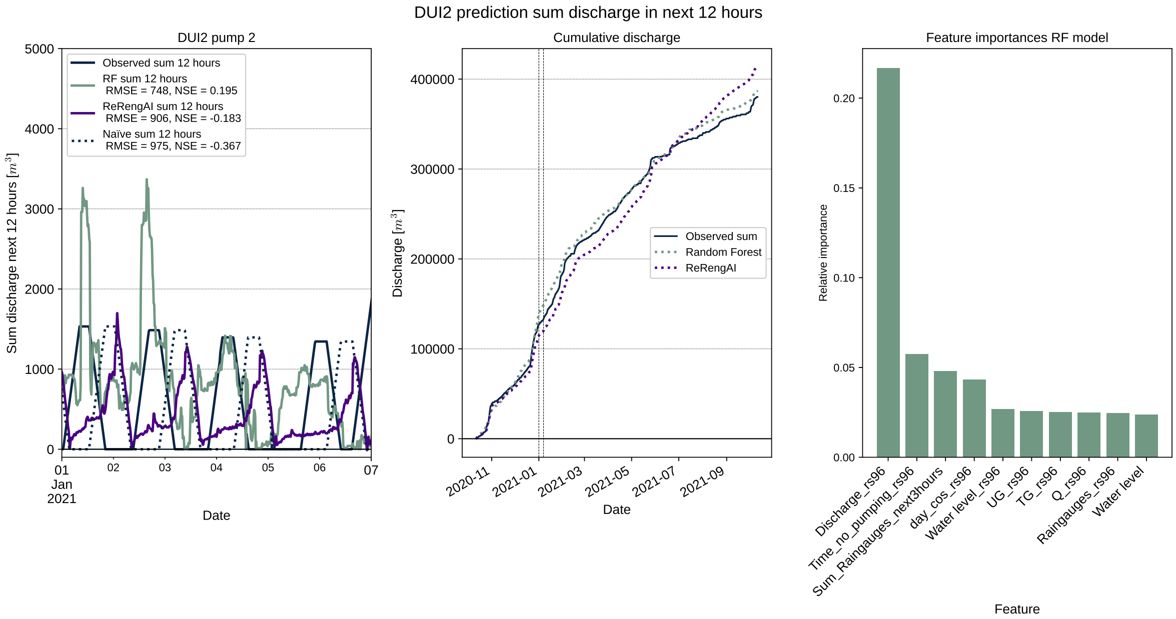


Figure A.14: Comparison Random Forest, ReRengAI, Naïve, and LSTM model for pump 2 of the Duifpolder (DUI2)

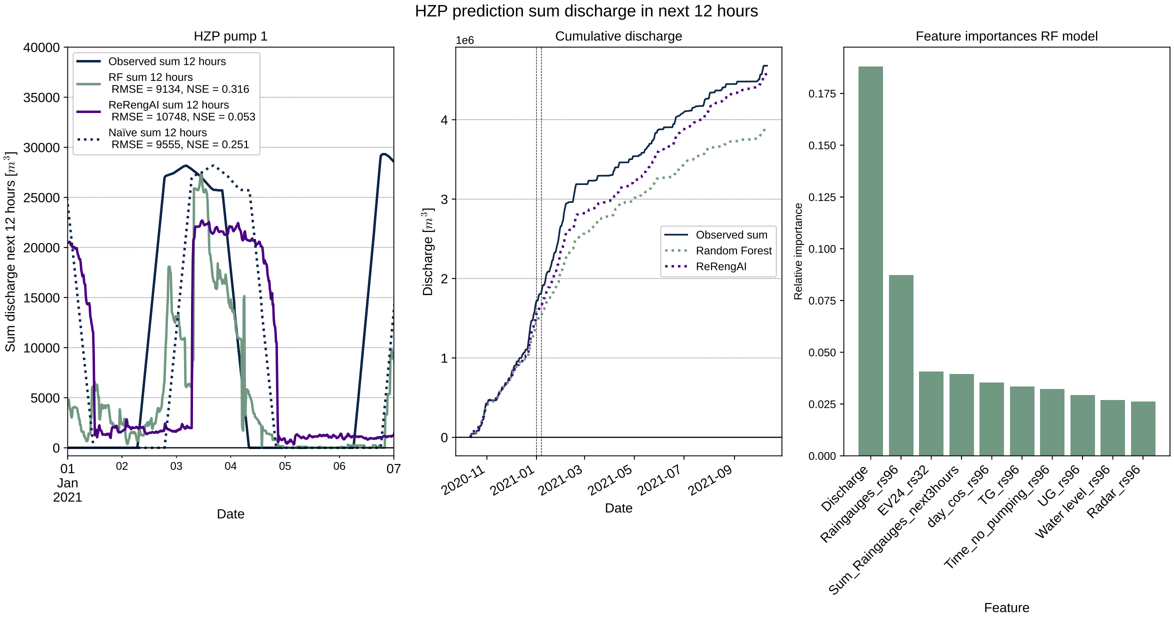


Figure A.15: Comparison Random Forest, ReRengAI, Naïve and LSTM model for the HZP

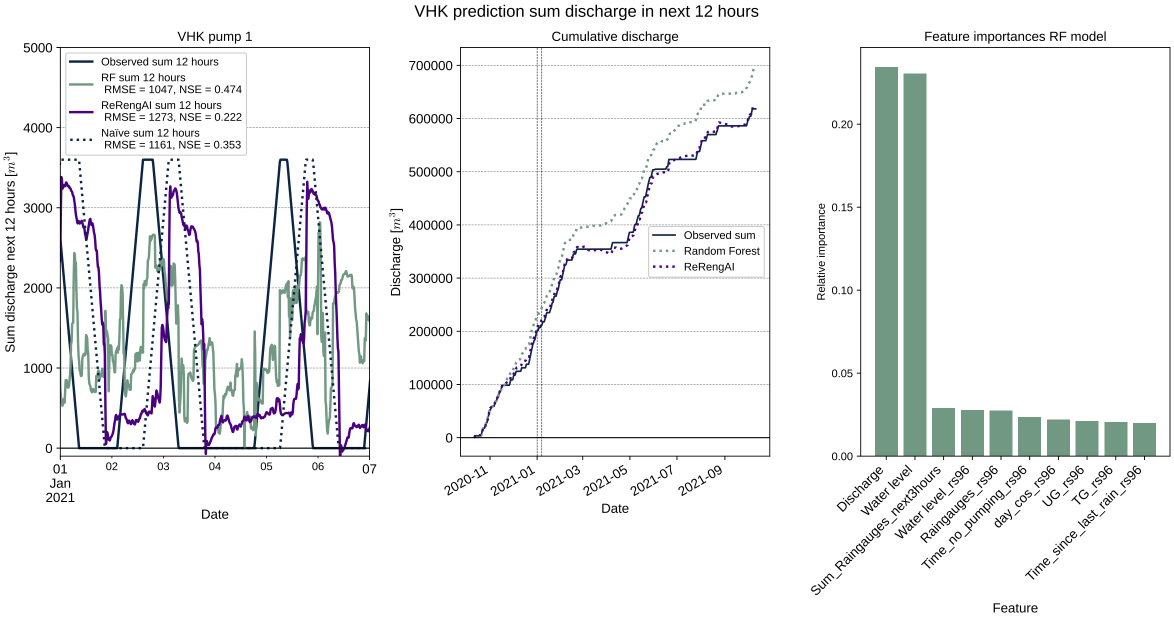
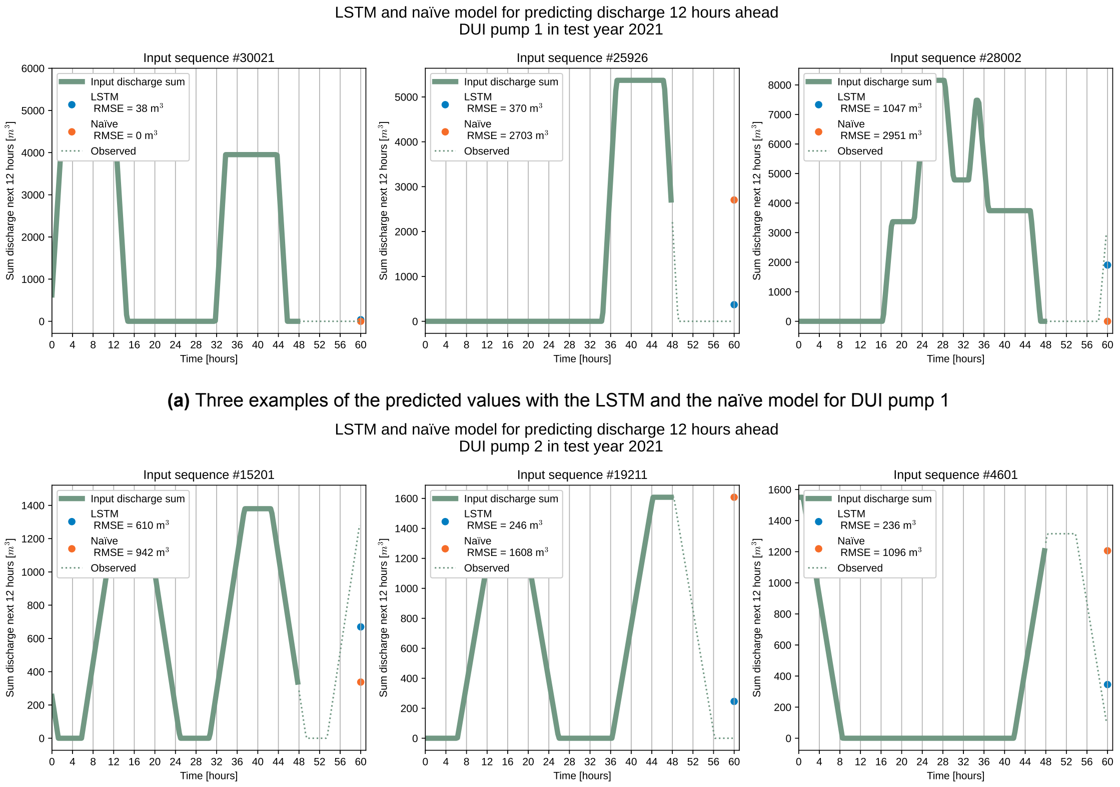


Figure A.16: Comparison Random Forest, ReRengAI, Naïve, and LSTM model for the VHK

A.5.3. LSTM model

The LSTM model is used to predict the sum of the discharges of the pump. In Figure A.17 some input sequences and outputs of the two different models are presented. Since this Figure only presents part of the predictions, the errors are calculated for all the predictions and presented in Figure 4.24.



(a) Three examples of the predicted values with the LSTM and the naïve model for DUI pump 1

(b) Three examples of the predicted values with the LSTM and the naïve model for DUI pump 2

Figure A.17: Plot of three random input sequences (sum observed discharge next 12 hours) with a length of 48 hours and the corresponding observed values and predictions of the LSTM and naïve model of the pumped discharge in the next 12 hours

A.5.4. Evaluation results

Boxplots and violin plots

Boxplots provide a concise summary of a dataset by a visual presentation of five numbers, the two quartiles (Q1 and Q3), the median (Q2) and the minimum and maximum value. The lower part of the box represents the first quartile (Q1), or 25th percentile, while the upper part represents the third quartile (Q3), or 75th percentile. The line inside the box represents the median. Whiskers extend from Q1 and Q3 to show the minimum and maximum values. Q1 and Q3 are medians of the lower and upper halves, respectively. The interquartile range (IQR) is the difference between Q3 and Q1. To determine the minimum and maximum, fences are calculated using Q3 + 1.5 × IQR for the upper fence and Q1 − 1.5 × IQR for the lower fence. The minimum is the smallest value above the lower fence, and the maximum is the largest value below the upper fence. Any value outside these fences is considered to be an outlier (Hu, 2020). Violin plots are an advanced version of boxplots and can provide additional insights in the data density. This is because it of the smoothed histogram of the data that is plotted along the y-axis.

A.6. Implementation

In this appendix some questions that have been discussed during the interviews at Delfland are presented. Next to that, the interviews that have been done in the period May - July are presented. The interviews have been done in Dutch and the main messages from the questions are presented below. The focus in the interviews was different for each person that has been interviewed. The main topics that have been discussed are the current system and the challenges in this operational system, the general view on deep learning and the prerequisites from Delfland for a good model.

A.6.1. Questions

The different questions that have been discussed during the interviews are presented in the list below.
Introduction:

- What is your role in the organization?
- What are the specifications of the current system?
- What are the challenges of the current system?

Deep learning:

- What do you know about artificial intelligence, and for which purposes do you use it?
- What is your view on artificial intelligence?
- Statements (Strongly disagree, disagree, neutral, agree, strongly agree)
    - AI has a positive influence on society.
    - In 10 years, a lot of jobs will have been taken over by AI.
    - AI may eventually surpass human intelligence.
    - The ethical issues raised by using AI are the biggest challenge.
    - I would trust a validated AI model to control the operation of the pumping stations.
    - A new model must show for at least 10 years that it has made the right decisions.
- What factors are considered the biggest disadvantages of deep learning? (put in order of big disadvantage to small disadvantage)
    1. Not yet enough knowledge about the models.
    2. Not known what will happen in extreme situations.
    3. Need a lot of computational power to make models.
    4. Training of models takes a lot of electricity.
    5. Black-box (difficult to interpret).
    6. Black-box (difficult to understand and explain).
    7. Depends on quality of input data.

Model:

- What is important for a good model? (put in order from important to less important)
    1. Safety of the system (risks).
    2. Insights into the uncertainty of the prediction.
    3. Good predictions in extreme situations.
    4. Good predictions in the daily situation.
    5. Quality of the prediction.
    6. Explainability of the model.
- Who is responsible when damage is caused due to a wrong choice of the model?
- What is needed to prove that a deep learning model can be trustworthy?
- How does the future DSS-system look like? (decision-supporting system)
- What are the developments in this field in Delfland or at other water authorities?

Constraints/conditions:

- Which time horizon is important for the prediction?
- Which irregularities in the data need to be taken into account? (changing summer to winter target water level, not measured inlets, irrigation in summer, quality of the data, etc.)

A.6.2. Interview 1

This interview was done with two representatives of the organization. One of them coordinates the water level operators and has been working at Delfland for many years. Another interviewee manages the DSS-system (BOS-system), and knows a lot about the model inputs and the current system.

General

Not a lot of buffer/storage in the system, which means that pumping too much for one day already means a low water level in the canals. The margin in the system is also very small +- 3 cm. This implies that the management is a challenge. We are satisfied with the current model, but there are improvements possible.

A challenge of the current system is the quality of the input data, this is not always 100% trustworthy and really influence the predictions and thus the operation. This is because weather forecasts that are used are expectations and it is very likely that these will change. Next to that, the control of the system is quite complex. Another challenge is that it is difficult to obtain insights in the decision-making of the model. In the extreme situations the water level operators take over control from the DSS system, so we are not completely dependent on the model.

The future of the system

The model will be a hybrid model in which the operation is controlled by a more advanced BOS-system, with data validation. In this BOS-system the RTC-tools model will be de-clustered, so that the inflow of the water from each polder can be calculated separately. Maybe also with a deep learning model, but still with people supervising the system. The system probably won't run fully automatically because there will be more extreme situations because of climate change. In the future water management of our system, several objectives will be optimized. We will not only be optimizing based on water quantity, but the system will be optimized also based on the water quality, energy consumption (minimizing greenhouse gasses, costs, etc.) and ecological purposes (fish migration, biodiversity, etc.). In the future we want to pump more energy efficient and take sustainability more into account.

Artificial Intelligence

"I have been listening to podcasts of AI for two weeks in a row now."

The people decide how AI will be implemented. For example if you compare nuclear energy with nuclear bombs, one purpose is serving humanity while the other destroys the world. The table below was filled in together to get an idea of the general view on deep learning of two representatives of the organization.

Table A.5: List of statements about AI

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
AI has a positive influence on society		x			
In 10 years, a lot of jobs will have been taken over by AI.					x
AI may eventually surpass human intelligence.				x	x
The ethical issues raised by using AI are the biggest challenge.					x ++
I would trust a validated AI model to control the operation of the pumping stations.			x		
A new model must show for at least 10 years that it has made the right decisions.		x			

We cannot imagine now what the impact of AI will be. It is up to the people to apply it in a way that it will help us. I think that it will definitely surpass human intelligence at some point. Yes, I think we can trust AI, but the quality of the data can be a problem.

Which factors are considered to be the biggest disadvantages of deep learning?

1. Depends on quality of input data.
2. Black-box (difficult to interpretate).
3. Black-box (difficult to understand and explain).
4. Not yet enough knowledge about the models.
5. Not known what will happen in extreme situations.
6. Need a lot of computational power to make models.
7. Training of models takes a lot of electricity.

Model requirements

What is important for a good model?

1. Good predictions in the daily situation.
2. Quality of the prediction.
3. Explainability of the model.
4. Safety of the system (risks).
5. Insights in the uncertainty of the prediction.
6. Good predictions in extreme situations.

"You need explainability to get the people involved."

Important considerations

- Use the model in parallel with the current prediction and check performance using historical data.
- Data is more important, so quality needs to be improved.
- We need people that have the knowledge about these type of models if we want to use it for operation.
- The model needs to be explainable to some extent, so that it is possible to back up decisions in order to convince people that the model did the right thing.
- Nature is way more powerful than we are. As a water board we try our best effort.
- The organization is accountable for damage.
- If a new pump is installed, it is difficult to implement it in the current model. Deep learning could be useful for this purpose.
- Depending on the weather the future predictions are important. In case of a lot of rain: predictions are more important.
- More extreme situations will happen in the future, so designing a model that can handle all these events will be a big challenge.

A.6.3. Interview 2

This interview is done with one of the people who currently coordinates innovation in the organization. This interviewee has worked for a very long time for the water authority of Delfland.

Introduction innovation coordinator

I am working on coordinating the innovation within the organization. For example looking at the steps in the innovation procedure and finding out which steps need to be taken in order to implement an innovation. Innovation in Delfland is about developing and applying new ideas that add value to Delfland and the surrounding. This is essential to tackle problems associated with climate change and the increase in population. Next to that, innovation is needed to keep up with the digitalization and the transition to a circular economy (Hoogheemraadschap Delfland, 2017).

Artificial Intelligence

I think that we as an organization should experiment with these new technologies, but always taking into account the ethics associated with Artificial Intelligence. AI can be of great value if you can adapt quickly and you discuss the ways in which you are using it. I am interested in the use of AI for the operation of the system as long as the technological developments are in balance with the developments on the ethical aspects along with the responsibility. For example, the considerations and reasons why a certain decision is made, need to be traceable.

A.6.4. Interview 3

Next to that, a business consultant in the data processing department has been interviewed.

Introduction role of business consultant at the water authority

We operate on the intersection of business and information supply. The information supply consists of ICT (information and communication technology) and IMG (information management and 'gegevens beheer'). We try as business consultant to find the functional question and to bring the different stakeholders together, for example in the field of information protection and privacy. Technical architecture is important, for example multi-factor authentication. The objective is to combine all the conditions and decide whether the request is relevant. Currently we are trying to make the processes in the organization more streamlined. In the past, we mostly did what we thought was right, until we arrived at the e-locket, together with ICT, architect, security officer, privacy officer, etc.

Companies state: 'Data, we have to do something with it.'

Now we try to change this by looking into the steps that need to be followed for bringing a question from the 'business' to a plan that can be executed and by consulting the different disciplines. The business in this case are the people that work with the data in for example the operational water management. So we are setting up a plan and finding the needs to come to the last stadium in the e-locket. The first step is set up a simple architecture check, here the main objective will be discussed. This step is done together with an architect and we discuss if it is feasible or not. If this does not work, a new plan is made. If it works, the next step would be to set up a project start architecture, in which all the details will be discussed. These processes we try to summarize, also together with the ICT-department. With as objective to have a detailed list of steps that need to be executed before a certain question from the business is translated to a specific action. In the implementation phase, we as business consultants move to the background and only check the result every once in a while.

As business consultants you have to ask questions like, 'Why do you need this certain application', 'Will this be used for operation, or only for insight?', 'Why do you need these insights and what are you going to do with this?' and more questions also in the field of privacy legislation and data protection. From a business perspective sharing data is not a problem, while looking from data protection point of view this can be confidential because this increase the vulnerability of us as organization.

In practice it can be observed that the gap between the programmers and the people working on the operation is big. It is difficult for the programmers to describe the possibilities of their expertise and to explain what they are doing, while for the 'business' people there is limited knowledge on these possibilities itself, which makes it hard to translate a demand to a feasible and efficient solution. In Delfland there is a program, 'data-driven' work, in which I am involved. This is necessary to make this 'translation' and on top of that to look into the construction of several dashboards in the organization.

Collaboration between and at the water authorities

As business consultants, we have little contact with other water authorities, since we work mainly internally. The 'business' does work together a lot. Every water authority has their own challenges, some are located near the sea and other water authorities have to deal with high and low discharges of several rivers. The operational water management in the Netherlands is complicated since some municipalities have to collaborate with different water authorities, which also applies to the provinces and water authorities. The DEEP (Data science & Engineering Expert Program) project is a initiative of the 'Waterschapshuis' in the field of AI. DEEP has as an objective to expand the expertise on data-driven work by educating young data-engineers and scientists in the field of water management. This program focuses more on data management, digital transformation and partly on machine learning models, but the focus is not yet on deep learning models. The 'Waterschapshuis' is driven by questions from the water authorities and could also be interesting to involve in this type of research. The 'Unie van Waterschappen' represents the water authorities in the national and international arena. They advocate for the interests of the water authorities and promote knowledge exchange and collaboration.

At the water authority a lot of departments work separately from each other, which has been like this from the existence of the water authority. We try to increase the cooperation between the departments, because this could be improved. For example the two separate locations in Delft and Vlaardingen create a geographical distance between colleagues, which does not enhance the cooperation.

Future of the water system

I do not think that in the future the system will be completely automatic, especially because there will be more extreme events. Another point is that I think that people want that there are people involved in the water management of the area. The question what error type people would rather accept; a human or a model error, is a interesting one, on which I do not have one answer. On one hand, everyone can imagine that it is possible to make a error as a human, for a computer that might be more difficult. People like to point a finger to the person who is responsible, while this is more difficult for a DL model. Nature is always the most powerful and a model not a person can beat that. The question of accountability is a challenge, since the model developer, but also the user and the politics are involved and partly responsible for the performance of the model.

The explainability of any model is important, especially in case of a wrong prediction. It is important that these errors can be explained, in order to win back trust.

Another interesting point in the consideration of using human 'expert' knowledge or automatic control by a computer is the difference in 'behaviour' in case of a stressful situation. People are able to feel things, such as stress or have something like gut feeling. An AI system does not have this feeling of stress, which can be beneficial, but can also be a drawback in some cases. Although both the decision process of a person as well as a deep learning model is not always insightful, which is a similarity.

The representatives at the organization were all very interested in the possibilities of using machine learning for the operational control of the pumps and open to discuss their thoughts, which provided valueble insights for this research.