Master's thesis – master Earth Surface and Water

# IMPROVING GROUNDWATER LEVEL MODELS WITH MACHINE LEARNING

*[Investigating the influence and significance of machine learning model selection and input datasets in order to improve the performance and prediction skill of groundwater level models that can be used to identify possible future low-flow periods and droughts]*


**Utrecht University**

Faculty of Geosciences

Earth Surface and Water MSc. Programme


Universiteit Utrecht

*June 2022*


*Author*

Balazs Bischof (6510241)

b.bischof@students.uu.nl

*Supervisors*

**First Supervisor:** Dr. ir. Niko Wanders

**Second Supervisor:** Sandra Hauswirth MSc


## Abstract

To respond to climate change and urbanization, water management systems will need to adapt in the next decades all over the world, including the Netherlands. Hydrological modelling and the simplification of real-world processes are vital for managing water resources and systems. In the future decades, machine learning (ML), deep learning (DL), and neural networks (NN) are projected to be critical in supporting humans in handling increasing volumes and diversity of data, extracting relevant information for a specific variable, and offering viable answers to crucial issues. Numerous articles have showed over the last decade that ML can help hydrologists to model transdisciplinary and complex systems that are challenging to simulate using standard numerical modelling methods. Machine learning and neural networks are becoming essential tools for hydrological analysis since they allow us to handle large amounts of data and extract significant and hidden information, as well as correlations between hydrological variables. The objective of this study is to enhance the performance and prediction skill of an existing groundwater level model by evaluating the impact and relevance of ML model selection and input datasets. For this purpose, a process-based ML approach was implemented, using the National Hydrological Model for physical consistency along with different types of input features including meteorological, hydrological, and environmental variables. The findings reveal that both applied methods are capable of predicting groundwater levels and boosting the numerical model's capabilities. To better represent and visualize these results a groundwater map was created for average summer conditions in 250m resolution for the whole area of the Netherlands. Furthermore, in order to facilitate future groundwater management and research, the feature importance was evaluated in various situations to examine the overall picture of variable relevance. The estimated feature importance values and the model's error results were further examined to determine whether there are any spatial pattern or trend in the outcomes. From this, it can be concluded that the model is suitable for modelling typical groundwater levels, but it suffers from significant error when predicting groundwater extremes, despite the fact that the errors and results are still more closely related to actual groundwater levels than the numerical model's results. As a result, the approach works poorer in the southern areas of the Netherlands, such as Limburg and Maastricht. Additionally, a model was also conducted to explore if the difference between the numerical model's outputs and real groundwater levels could be estimated. Different scenarios were investigated, and a generic, simplified model was developed which can predict the errors between simulation values and actual groundwater observations with an adequate accuracy. This simplified model might help to model hydrological and environmental processes, since by using this model groundwater level predictions can be generated without knowing any actual groundwater level values. In summary, the work includes a detailed description of methodology, demonstrating the required steps in creating a machine learning model that can predict hydrological processes. The findings can be used in future study to improve groundwater level predictions and, as a result, water management strategies in order to reduce the detrimental effects of future groundwater level extremes that could result in severe droughts or floods.

# Contents

# 1. Introduction

Groundwater constitutes for approximately 30% of the world's total freshwater (including ice and snow as a freshwater source), thereby is one of the most important components of the hydrological cycle with a wide range of socioeconomic and environmental implications, including infrastructural security, food production and ecosystem sustainability (Gleeson et al., 2016). Groundwater reserves are critical for more than 7 billion people who depend on them for drinking water, agriculture, and industry (Wada et al., 2014). It is also a significant freshwater supply for domestic and industrial water usage and accounting for over 38% of global irrigation demand (Siebert et al., 2010). However, in recent decades, these essential water sources have been utilized inappropriately in many regions around the world (FAO, 2021). As a result of poor water management and increasing population (which implies an enhanced water demand) groundwater aquifers are being exhausted. Unsustainable extraction rates are surpassing recharge rates, resulting in long-term depletion of aquifers around the world (Bierkens et al., 2019). The depletion of aquifers has several negative consequences, such as greater energy costs for pumping water from deeper wells, land subsidence, lower river flow rates and deteriorated water quality (Alley et al., 1999). Furthermore, future climate change could have a severe impact on groundwater recharge and storage, exacerbating this unsustainable state. The influence of climate change on groundwater has just recently attracted attention, although subsurface water is one of the most important freshwater resources and critical for water and food security. In the future decades, the adaptation of water management practices, particularly decisions concerning the groundwater system, will be a crucial question for humanity.

Groundwater quantity and quality in the Netherlands has become an increasing problem in the last decades. The pressure on current groundwater resources is intensifying as a result of growing demands and droughts. Socioeconomic changes, climate change impacts, drinking water extraction, existing groundwater contamination, and increased usage of the subsoil, such as for aquifer thermal energy storage (ATES), geothermal energy, or mining activities, are all contributing to this pressure (Lijzen et al., 2014). Furthermore, future climate change consequences could exacerbate this stress. Precipitation shifts from summer to winter are projected, along with increasing evaporation demands in the summer due to warmer temperatures, which can possibly lead to longer and more intense dry spells (Philip et al., 2020). As an example, many water managers saw the drought of 2018-2019 as a wake-up call in a country that has typically focused on getting rid of water surpluses (Brakkee et al., 2021). In addition to climate change, variations in land use have an impact on groundwater recharge and levels. Since the sustainable management of these groundwater aquifers is critical to assure freshwater supply for all sectors, it is important to assess the consequences of climate change and land use change on groundwater aquifers (Brakkee et al., 2021). The mentioned threats (extreme droughts caused by the changing climate patterns or the excessive groundwater pumping rates) are placing pressure on national and regional water managers to develop suitable long-term plans for investments and adaptive measures that will lead to long-term water management that is both sustainable and robust (Verkaik et al., 2021). In order to solve societal concerns related to mitigation and adaptation to climate change, as well as strengthening climate resilience in general, detailed knowledge of the water table representing the groundwater system is necessary (Koch et al., 2021). Accurate, realistic estimates of groundwater levels are required to assist environmental decision-making that addresses current and future challenges. These estimates can help in the long-term management of groundwater and the prevention of negative outcomes (e.g., consequences in the agricultural sector due to severe low-flow periods, droughts, and floods).

Water management techniques will need to alter in the future decades across the world, including in the Netherlands, to react to climate change and urbanization. In order to manage water resources, hydrological modelling and the simplification of real-world processes are essential.

Historically, mainly process-based numerical, multi-physics simulation models such as MODFLOW have been used to evaluate groundwater depths and other water budget elements (e.g., runoff or soil moisture). The complexity of these hydrological processes and the associated computational demand makes these process-based modelling method challenging for operational applications. As a result, the spatial scales and accuracy that are required for adequate decision-making can typically not be provided by numerical, physically based models alone (Koch et al., 2019). This constraint is mostly due to the high computational demands of such models, which prevents detailed conduct calibration, sensitivity, and uncertainty analysis at high resolution (Asher et al., 2015; Stisen et al., 2018). With the advancement of sensors and satellites hydrologists now have access to far more data than in prior years. Given the drawbacks of physics (or process)-based modelling, as well as the increased availability and volume of environmental data, new opportunities in the modelling of various natural processes are emerging. Nearing et al. (2020) proposed that large-scale hydrological data may include substantially more information than hydrologists have been able to interpret into theory or process-based models. With recent improvements in machine learning (ML) and artificial intelligence, the performance of models based on a thorough understanding of physical processes can be enhanced. ML is a collection of tools that enables us to create and train models to extract and recreate spatial and temporal characteristics in datasets (Shen et al., 2021). The primary concept of ML and deep learning is to minimize human interference in feature creation and to promote maximum information extraction from data (Goodfellow et al., 2016). ML enables high-resolution modelling of water table depths that exceeds the spatial resolution and overall accuracy of traditional numerical physically based hydrological models (Koch et al., 2021). These algorithms identify patterns in datasets and use these discoveries to forecast future events. The approximate behaviour of a complex system (in this research focusing on processes linked to groundwater dynamics) may be represented using multiple ML applications and big data, with the potential of providing accurate predictions at a reasonable cost. Consequently, according to Sahoo (2017), data-driven and ML approaches based on nonlinear interdependencies can estimate groundwater level change without a detailed understanding of the underlying physical parameters. Although these ML applications are intriguing, they lack process descriptions, which prevent trained models from making predictions outside the training dataset's observed ranges (Koch et al., 2021). According to Reichstein et al. (2019), ML will become more prevalent, by combining ML and numerical models, it can help advance present modelling systems. This knowledge-guided ML (or hybrid modelling) technique aims to increase model performance and robustness by incorporating physical consistency into ML algorithms (Koch et al., 2021). New and large-scale interdependencies can be identified using a combination of ML learning and numerical models, which might help us better comprehend complex natural processes.

It can be expected that, in the upcoming decades ML, deep learning (DL) and neural networks will be essential in assisting us managing increasing volume and diversity of data, extracting meaningful information for a particular variable, and presenting potential answers for complex questions. In recent years, these modelling techniques have received a lot of attention from the water science and hydrology communities. Recent developments have enabled completely data-driven approaches to estimate groundwater levels with different ML techniques (Koch et al., 2021; Hauswirth et al., 2021; Sahoo et al., 2017) and artificial neural networks (Banerjee et al., 2011). As an example, in order to estimate the uppermost water table depth in average summer and winter circumstances throughout the full region of Denmark, Koch et al. (2021) employed a knowledge-guided gradient boosting decision tree model. The research revealed that by utilizing knowledge (physics) driven ML techniques, it is possible to precisely estimate groundwater levels with exceptional spatial accuracy. By analysing the sensitivity of the MLP (multi-layer perceptron) neural network, Sahu et al. (2020) demonstrated the relevance of feature selection (for input variables such as groundwater levels, precipitation, temperature, and river flow). They examined the sensitivity of different feature selections in three distinct sites in California, USA and came up with training dataset suggestions to get more accurate predictions. They showed that precipitation and river flow are relevant characteristics in many, but not all locations, nonetheless, creating reliable forecasts using only

temperature and historical groundwater level data is insufficient. The findings proved that while analysing the relevance of various input factors, it is necessary to account for location dependency. Climate, groundwater extraction, and surface water flows all have complex relationships with groundwater level in agricultural regions. Sahoo et al. (2017) developed a modelling framework based on spectral analysis, ML, and uncertainty analysis to gain a better understanding of the respective relevance of each factor and to estimate changes in groundwater levels. They proposed that this modelling framework may be used to simulate groundwater level change and water availability as an alternative to traditional methods, particularly in areas where subsurface parameters are unknown. Kraft et al. (2020) developed a new hybrid modelling technique that learns and predicts global spatio-temporal variations of observable and unknown hydrological variables. The results demonstrated that the model accurately reproduces the observed water cycle variables (evapotranspiration, runoff, snow water equivalent, and variations in terrestrial water storage). Furthermore, many ML techniques were applied to investigate different hydrological processes, such as using long-short term memory (LSTM) networks for discharge predictions in ungauged basins with rainfall-runoff data from catchments in the United States (Kratzert et al., 2019), employing advanced ML approaches to address difficulties in the mitigation of urban water hazards (Allen-Dumas et al., 2021), estimating groundwater nitrate concentrations on a large-scale level using several ML techniques (e.g., multiple linear regression, random forests and boosted regression trees) (Knoll et al., 2019), implementing an LSTM neural network to construct an integrated framework in order to estimate the snow water equivalent (SWE) based on daily snow observations (Meyal et al., 2020), proposing ML based algorithms for global design flood predictions (Zhao et al., 2021) or introducing novel ML models in order to map the susceptibility of the erosion of soil (Mosavi et al., 2020). Artificial intelligence techniques to predict groundwater levels and to model different processes connected hydrological sciences are becoming more popular and gaining the attention of many academics in the field. ML can assist hydrologists in modelling transdisciplinary and complex systems that are difficult to simulate using traditional numerical modelling approaches. ML and neural networks will be critical tools for hydrological analysis in the upcoming years as they allow us to handle massive volumes of data and extract meaningful and hidden information, as well as relationships between hydrological variables.

It is expected that climate change, urbanization and population growth will increase water demand and consumption. Therefore, to preserve current agricultural practices and increase water security, the development of more realistic and accurate hydrological models is particularly important. The modelling of these complex hydrological processes is challenging. Process-based numerical models are expensive, time-consuming, and most importantly cannot adequately represent hydrological processes with the required spatial scale and accuracy. Thus, the development of modelling tools and the implementation of innovative modelling frameworks are necessary. Due to the absence of large amounts of data in recent decades, employing ML approaches to model the groundwater domain has not been a frequent topic for hydrological research. However, recent droughts and low-flow episodes have highlighted the importance of having effective groundwater models and projections. National water authorities (for example Rijkswaterstaat in the Netherlands), who are responsible for the sustainable and safe management of water, must update their water management technologies in order to increase drought preparedness and accurately predict extreme groundwater levels in the upcoming future. Accurate short- and long-term predictions of groundwater levels would aid hydrologists in capturing high and low flow events, hence improving water management planning and mitigation approaches, with huge agricultural implications and major improvements in urban water management.

The main objective of this study is to show a potential of a groundwater ML model to estimate the depth of groundwater for the entire area of the Netherlands (41800 $km^2$). During the modelling process ML and big data analysis have been utilized to improve current modelling approaches and groundwater simulations. This was accomplished by creating two separate machine learning models

(Random Forest and CatBoost). The selection of input features for these models (meteorological, physical, and topographical) were based on their possible impact on groundwater levels. Additionally, models were introduced to new training datasets (e.g., land-use type, vegetation-index, soil type) next to the original training datasets. The purpose of introducing additional input data is to verify the importance of different parameters and assess their impact on model performance and predictions skills. A physically based numerical model's outputs were also used as an input variable. Potentially, implementing this physically based ML approach, the performance of ML models will be improved by coupling physical processes, allowing physical laws, such as mass balance or the Darcy's law to connect different model components in a more realistic way. This hybrid-modelling technique could better explain and investigate complex natural systems than process-based or data-driven approaches alone, allowing for a greater understanding of these processes. The research focuses on the possibility of improving the existing numerical model to demonstrate the potential of machine learning for accurate groundwater level modelling. In order to enhance the performance and prediction skill of groundwater level models, the influence and relevance of ML model and input dataset selection are examined. To show the potential for improvement, a thorough comparison of various ML model selection and input variable selection was conducted. A groundwater map was created in addition to the standard assessment metrics to illustrate the increased groundwater levels and the error between the improved ML model and the existing numerical model. Comprehensive assessments of current groundwater levels have been conducted in recent decades using observations and process-based numerical models. However, no systematic study focusing on applying machine learning techniques to estimate present (and likely future) groundwater levels over the whole Netherlands has yet been published. Since the groundwater system plays a significant role in developing adaptation and mitigation approaches, as increasing groundwater levels enhance flooding and decreasing groundwater levels exacerbate droughts, such a model might support environmental decision-making. Koch et al. (2021), Hauswirth et al. (2021) and Wang et al. (2018) demonstrated that utilizing decision trees and random forest regression to reliably estimate groundwater levels is a potential approach. Based on the findings of these research, the focus of this paper's modelling is on ML techniques and methodologies that use decision trees and random forests. The projected result of this research will identify the potential of ML models, with adequate setup and input variables, to forecast future drought spells and assist water scientists in adapting current water management techniques. The produced map can potentially be used as a starting point for more detailed groundwater models in the future.

To summarize, the primary goal of this work is to investigate the influence and significance of ML model selection and input datasets to improve the performance and prediction skill of groundwater level models that can be used to identify possible future low-flow periods or droughts. The findings can be applied for further research to enhance groundwater level predictions and, as a result, water management approaches in order to minimize the negative effects of possible future groundwater level extremes that could lead to severe droughts or floods. The study provides a thorough description of methodology, demonstrating the procedures involved in developing a ML model capable of predicting hydrological processes. The results are then discussed in depth, as well as their potential applicability in real life.

## 2. Methods

In this section the study area and the datasets used for training are introduced. The configuration of the ML model and variables are described, including the modelling framework,

implemented ML algorithms, supplementary datasets, and the properties of the employed physically based model. Furthermore, the different parameter tuning, assessment metrics, and sensitivity analysis methodologies employed during model construction are discussed.

## 2.1 Study area

This research is carried out for the entire Netherlands, located in Western Europe, and covering approximately 41,800 $km^2$. The Netherlands is mostly flat, with the highest point being 327 meters above sea level. Since about 26% of the country's land area is below sea level and more than half of it is extremely prone to floods, a thorough assessment of the Dutch groundwater system is critical. Additionally, according to Klein Tank et al. (2014), Brakkee et al. (2021) and Philip et al. (2020) changing precipitation patterns and evaporation demand due to climate change might lead to longer and more intense dry periods. The agricultural sector plays an important role in the countries' economy (54% of total land surface being used as a farmland). As groundwater is essential to meet crop water requirements, massive fluctuations without mitigation strategies can result in significant losses. As a result, the Netherlands, like many other nations throughout the world, requires precise, accurate, and realistic groundwater models to minimize and prevent the damaging effects of future groundwater level extremes.

## 2.2 Input variables

### 2.2.1 Meteorological variables

Input variables were selected based on their physical importance to the target variable (i.e.: groundwater levels). Various features were used during the development of the ML algorithm that are likely to have a direct or indirect impact on groundwater levels. The two main meteorological feature which plays an important role in the fluctuations of subsurface water levels are precipitation and evapotranspiration. Meteorological data was gathered for the 13 major weather stations in the Netherlands (*Fig 1.*) obtained from KNMI. The various data values for different regions represent meteorological variability, which could have a significant impact on groundwater level dynamics. Daily precipitation and evapotranspiration observations are included in these files, which were transformed and summed to weekly data. The quality of these datasets was excellent, although missing values were corrected on some occasions based on weather seasonality.

### 2.2.2 Hydrological parameters

Hydrological parameters were also included to the ML model to improve its prediction ability using data from the Rijkswaterstaat. The primary water flows in the nation were represented by discharge measurements from the country's principal rivers and inflow locations (Rhine river at Lobith and Meuse river at Eijsden Gens) as well as seawater level observations (SWL) from Haringvliet (*Fig 1.*).

### 2.2.3 Environmental variables

Land use and soil data were also employed to improve model performance and gain insight into whether these features have a significant impact on groundwater level prediction. Land use maps represent the spatial distribution of different physical properties of land coverage. The LGN4 (Landelijke Grondgebruikskartering 4) model was used, developed by the Wageningen Environmental Research in 2003. The LGN4 file is a 25-meter-resolution raster file that distinguishes 39 different forms of land use. The file categorizes the most major agricultural crops, as well as a number of natural and urban classifications. Satellite pictures from 1999 and 2000, as well as other related geographical data were used to build the file. In order to synchronize different datasets, the resolution of the model was changed to 250 meters. This equals the resolution of all the used data including the physically based groundwater simulation model. The 39 different land use classes were reduced to 16 categories to ensure that each land use is sufficiently represented in the training of the ML model (the original and changed classes can be found in *Appendix A.*). This data was implemented as a categorical data

type. *Appendix A* includes *Fig. A/1.* and *Tab. A/1.* which shows the land use map of the Netherlands and the distinct classes, respectively. Employed classes of land usage can be found in *Tab. A/2*.

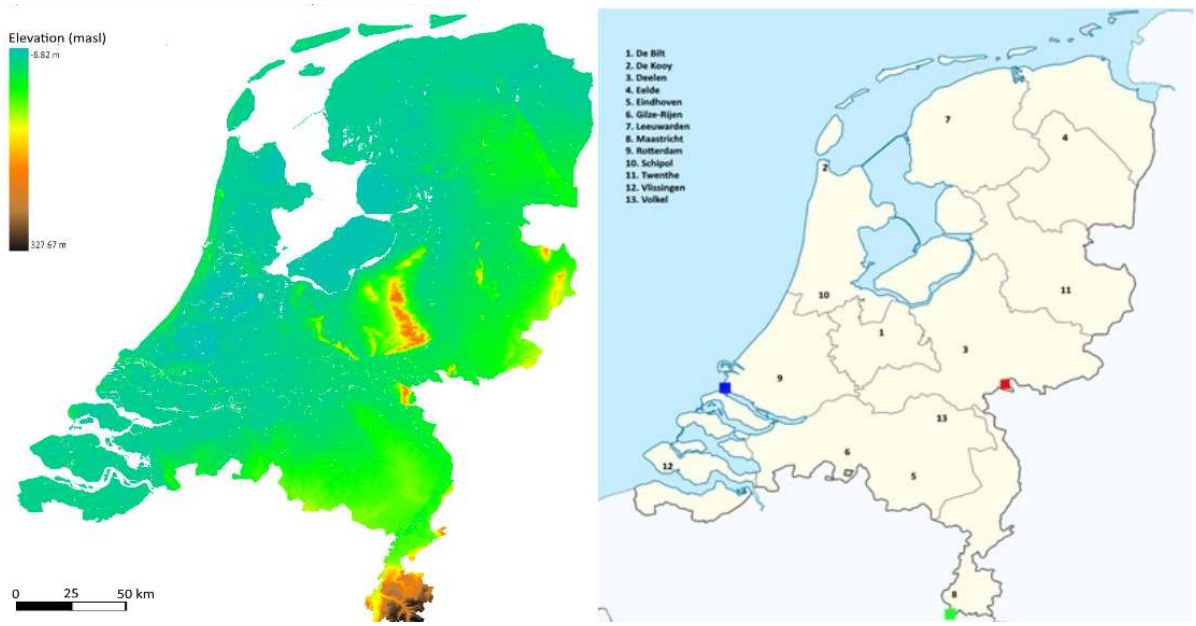
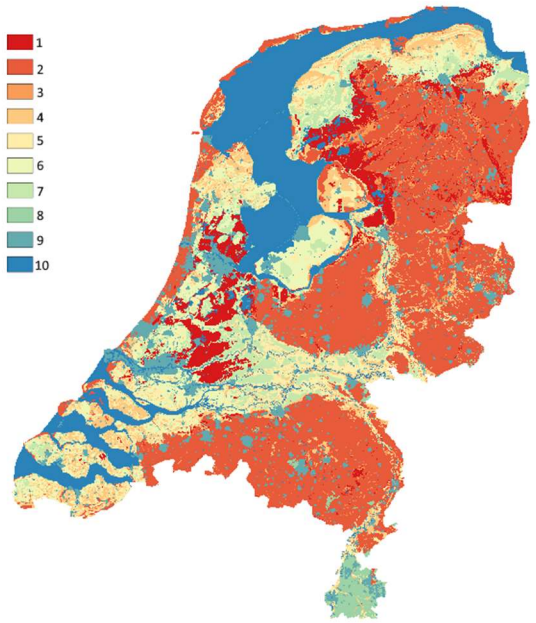



*Figure 1. Left: Digital Elevation Map of the Netherlands. Same data was used in the model as an input feature. Right: Regions with different meteorological observations and the main hydrological measurement locations. The numbers show the locations of meteorological observation stations. The rectangles show the sea-level measurements at Haringsvliet (blue), the discharge observations of the Rhine at Lobith (red) and the discharge observations of the Meuse at Eijsden Grens (green).*




| Category | Soil type |
|---|---|
| 1 | Peat |
| 2 | Sand |
| 3 | Swampy sand |
| 4 | Light loam |
| 5 | Heavy loam |
| 6 | Light clay |
| 7 | Heavy clay |
| 8 | Loam |
| 9 | Buildings |
| 10 | Water |

*Table 1. Soil type categories in the used dataset*

*Figure 2. Soil type map of the Netherlands*

Additionally, the distance between each cell and the nearest water body (canals, ponds, rivers, lakes, and the North Sea) was calculated and added as an extra input feature to check for a potential link with groundwater levels. The same land use dataset and map were used in this computation, with a resolution of 250 meters. The calculations were made in meters.

Furthermore, a soil type map (*Fig. 2.*) was employed as an additional input feature by using the soil type map (Grondsoortenkaart 2006) of Wageningen Environmental Research from 2006. The soil type map shows the location of the peat soils and mineral soils in the Netherlands. The various soil types have been encoded and 10 categorical variables were used in the model (*Tab. 1.*). Elevation data was also used as an input feature to see whether there was a relationship between groundwater level and topography. Data from AHN (Actueel Hoogtebestand Nederland) used with a resolution of 250 meters. *Fig. 1*. shows the visualization of the implemented dataset.

### 2.2.4 Groundwater simulations

Additionally, the modelling methodology was based on a hybrid approach, thereby the results of a physically based groundwater simulation were employed as a supplementary dataset alongside the groundwater level observations. With such a model and additional data, it is feasible to increase the models' performance and robustness by incorporating physical consistency in the original, data-based model. The National Hydrological Model was used, which is the integrated country-wide ground and surface water model of the Netherlands, developed by Deltares and WENR. The groundwater simulations were performed with a spatial resolution of 250 meters and over a time span of 1980 to 2019, which mostly corresponded to the time frame of the data utilized as an input variable. Firstly, the simulation results were combined with the groundwater observation dataset. Groundwater level values from the simulation results were selected for the related coordinate (with a total of 4002 well locations) over the investigated period with weekly temporal resolution. The original input dataset (which included meteorological and hydrological measurements and observations) was then merged with the groundwater observation and simulation values on the appropriate date and coordinate values by region. *Tab. 2.* includes the summary of all the utilized input features, including basic information and the source it was obtained from.

| Variable | Description | Source |
|---|---|---|
| **Precipitation** | *Daily precipitation data for 13 separate regions nationwide, summed into weekly data for modelling* | KNMI (Koninklijk Nederlands Meteorologisch Instituut) |
| **Evapotranspiration** | *Daily precipitation data for 13 separate regions nationwide, summed into weekly data for modelling* | KNMI (Koninklijk Nederlands Meteorologisch Instituut) |
| **Discharge of Rhine** | *Weekly discharge measurements from Lobith* | Rijkswaterstaat Waterinfo |
| **Discharge of Meuse** | *Weekly discharge measurements from Eijsden Grens* | Rijkswaterstaat Waterinfo |
| **Sea-water level** | *Weekly seawater level observations from Haringsvliet* | Rijkswaterstaat Waterinfo |
| **Groundwater simulations** | *Groundwater simulations for the period 1980 to 2019 with a spatial resolution of 250 meters generated for the whole Netherlands* | NHM (National Hydrological Model), (Deltares, WENR) |
| **Land use type** | *Land use map of the Netherlands from satellite observations in 2003. The original resolution of 25 meters and classes of 39 were changed to 250 meters and 16 classes, respectively* | LGN4 (Landelijke Grondgebruikskartering 4), (Wageningen Environmental Research) |

| | | |
|---|---|---|
| **Soil type** | *Soil type map from 2006 including 10 different categories* | Grondsoortenkaart 2006 (Wageningen Environmental Research) |
| **Elevation** | *DEM (digital elevation map) of the Netherlands. In order to keep the model consistent, the original 100-meter resolution was changed to 250 meters* | AHN (Actueel Hoogtebestand Nederland) |
| **Water distance** | *Water distance was calculated for every cell (pixel of 250-meter resolution) using the aforementioned land use data* | Calculated (base model was LGN4) |

*Table 2. Summary of the used input features*

## 2.3 Target variable

Groundwater level data has been collected from 1980 to 2019 (39 years) for 4002 observation wells all located within the border of the Netherlands. The data was obtained from DINOloket, which hosts publicly available subsurface data from the Netherlands Geological Survey (TNO) and the BRO (Basisregistratie Ondergrond). Although the utilized dataset is of high quality, the registered periods are sometimes different and shorter than the original time frame. Since the groundwater level data is deficient considering extreme values (very low or high groundwater levels) the 5% tails of the dataset were eliminated, in order to increase the performance of the model.

## 2.4 Modelling framework

### 2.4.1 Data pre-processing

New potential in the modelling of many natural processes is arising as a result of the increased availability and volume of environmental data. With sufficiently lengthy and detailed datasets, data-driven modelling approaches (in this instance, machine learning) may be used to make accurate hydrological forecasts. For several hydrological and meteorological variables, the Netherlands possesses outstandingly long and high-quality observational records. In this study, various forms of hydrological, physical, and meteorological observations, such as discharge measurements from major rivers, soil type conditions, land use features, distance from water bodies, evaporation, and rainfall data, were gathered over the period 1980 to 2019.

Even though data in the Netherlands is typically of good standard, real-world data contains errors, noise, partial information, and missing values. The ML algorithm must be able to quickly comprehend the data's attributes to be accurate and exact in predictions. Therefore, the datasets quality has been increased by cleaning missing data, either by eliminating it or through imputation (either with best guess or estimating it by considering seasonality). Outliers may have a negative impact on the performance of our machine learning method; thus, these values have been identified and removed. A general overview and visualization of the used datasets can be found in *Appendix B.*

### 2.4.2 Training and testing

The data was split into 50% training and 50% testing manually, in order to fully separate distinct well locations. To provide an unbiased estimate of the test set error, random forests do not require cross-validation or a separate test set. During the run, it is approximated internally. This division allows the model to be validated for nearly 2000 distinct well locations, demonstrating the generalization's effectiveness. To show the performance quality of the investigated ML algorithms the results were compared with the evaluation metrics of the numerical results. The used parameter settings (results of the hyperparameter tuning) can be found in *section 2.4.5.* The model is calibrated for the 90% of the existing data, since in the tails of the dataset (5% smallest and 5% largest groundwater values) does not contain sufficient amount of extremely high or low values, thereby it

has a detrimental impact on the model's performance. Due to this constraint the model is not being trained on extremely high (primarily occurring in the southern Netherlands, particularly Tilburg and Maastricht) and extremely low groundwater level data (mostly typical in the northern regions). As a result, the model's efficacy in predicting groundwater level extremes is restricted.

The research examines the effects of model selection, with one of the most anticipated outcomes being the capturing of correlations between groundwater levels and other dependent elements including meteorological and hydrological observations and a process-based national groundwater model of the Netherlands. The two implemented machine learning methods were compared to see how well they improved the numerical model outcomes. By using different evaluation metrics and visualization practices the generalized groundwater model have been validated, thus a nation-wide groundwater level map could be developed. Furthermore, a thorough analysis of the spatial differences in model performance and regional variations in feature importance was carried out using the spatial separation of the model's outputs. In addition, a separate machine learning model (using the same ML methods) was applied to estimate the differences between the physically based numerical model outputs and the actual groundwater level. As a result, next to the country's groundwater level map, an error map was created, which may be utilized as a starting point for future groundwater models and research in general. *Fig. 3.* visualizes the general overview of the implemented modelling framework.

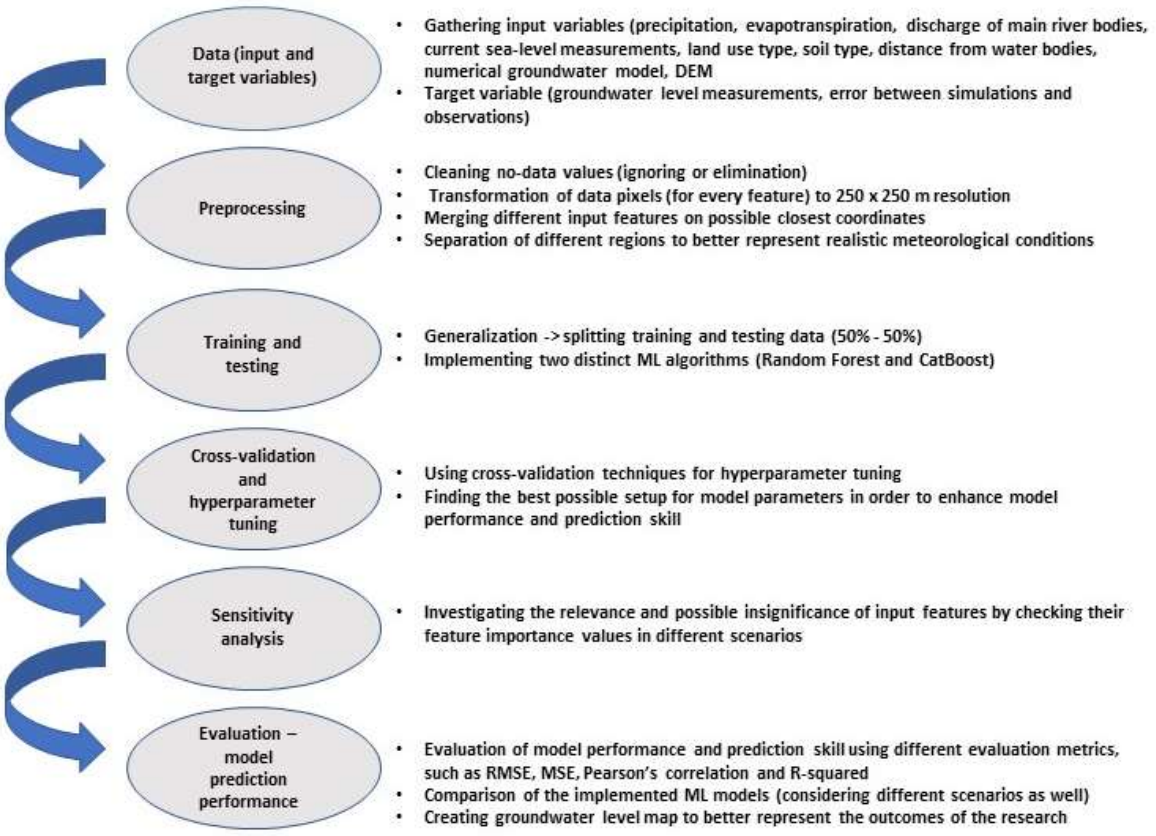



*Figure 3. Methodological framework of the implemented groundwater model*

### 2.4.3 Random Forests

Two distinct machine learning approaches were used in this study. Random Forests (RF) and Gradient Boosted Regression Trees, as noted in *Section 1.*, are widely employed in the hydrological sector, owing to their relative ease of implementation and the excellent quality of outcomes. These

methods have been tested and evaluated to see how well they function in terms of predicting groundwater levels.

RF are an ensemble learning approach for classification, regression, and other problems. Given its outstanding or great performance across a wide range of classification and regression predictive modelling tasks, it is arguably the most common and widely used machine learning method. The underlying idea of this approach is that many uncorrelated models are working together to make decisions that will outperform any single model (i.e., Decision Trees alone). RF are consisting of several Decision Trees, which are a sort of regression model that is built in a tree structure, with the data being split into subsets many times until no more can be created (*Fig. 4.*). The predictions from the trees are averaged over all Decision Trees in the model, yielding better results than any single tree. The ensemble's models are then utilized to make a prediction for a new sample, and the forecasts are averaged to provide the forest's prediction (Applied Predictive Modelling, 2013).

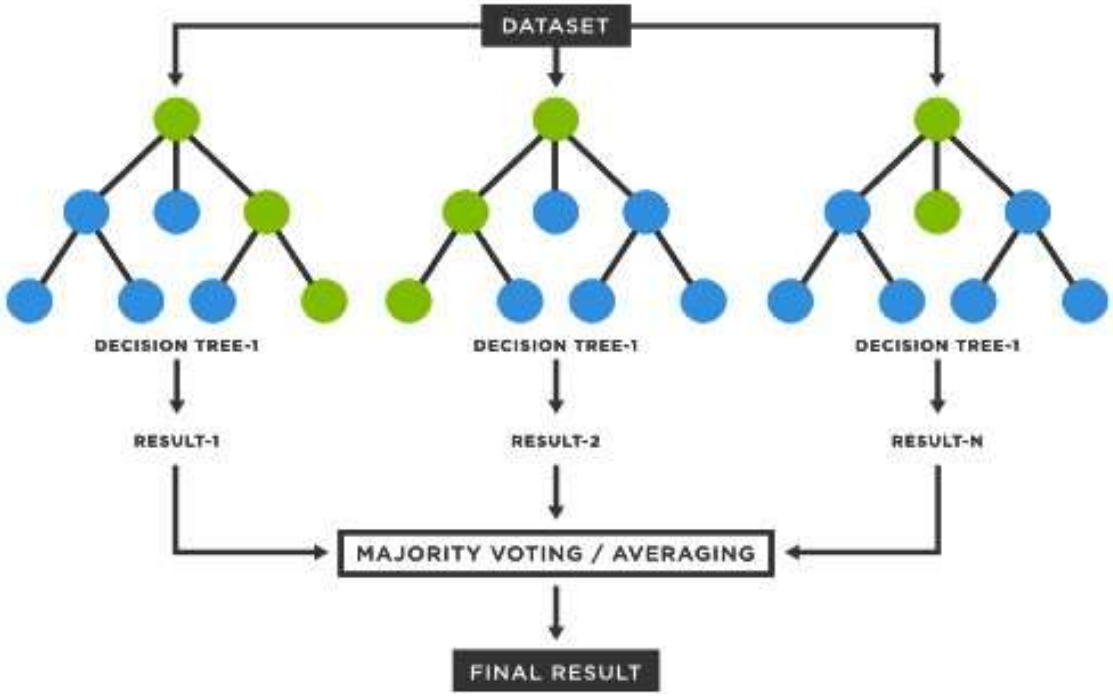



*Figure 4. Simplified framework of Random Forest models (source: What is a Random Forest? | TIBCO Software)*

### 2.4.4 CatBoost

In addition, a type of Gradient Boosting approach (i.e., CatBoost) was applied in this study to determine if it could outperform the RF algorithm's predictions. Gradient boosting refers to the process of enhancing a single weak model by merging it with several additional weak models to create a collectively strong model. It can detect any nonlinear relationship between the target data and the implemented features. CatBoost (CB) is an open-sourced machine learning algorithm developed by Yandex. This method was developed to create more generic models for massive datasets at an exceptional computation speed. The method was chosen over other Gradient Boosting Decision Tree algorithms (such as XGBoost) because it can function without a lot of data preparation and has significantly higher prediction skills.

### 2.4.5 Hyperparameter tuning

When it comes to model quality, obtaining more data and feature engineering usually pays off the most, nevertheless, when there are no other possibilities to gather more data, the model

performance can be further enhanced by hyperparameter tuning. The settings of an algorithm that may be modified to improve performance are known as hyperparameters. Calculations were performed for both developed ML algorithms utilizing different packages from the Scikit-learn library that use cross validation approaches to see which parameter setting results in the best performance and prediction skill. Using a K-fold cross validation approach, the Scikit Learn's RandomizedSearchCV was used with a wide variety of hyperparameter values to narrow down the available options. This approach selects combinations randomly and attempts to identify the greatest possible combination, making it computationally less expensive. The basic framework of such a cross validation method is the following:

1. Splitting the data into groups
2. Taking one group as test dataset
3. Training the model with the remaining groups
4. Fitting model on the test dataset
5. Evaluating the score and comparing it with the performance of other groups with different parameters

So, in conclusion, this method randomly divides the collection of observations into k groups, or folds, of roughly similar size. The first fold is used as a validation set, and the model is fitted on the remaining k - 1 folds (An Introduction to Statistical Learning, 2013). The results of the randomized search of different hyperparameters including a short description can be found in *Tab. 3.* and *Tab. 4.* for CB and RF, respectively.

Subsequently, given the previously filtered down results, Scikit Learn's built-in package GridSearchCV was used to determine the optimal parameter sets for the CB algorithm. The randomized search results were manually enlarged using the closest neighbouring values. Following that, GridSearchCV analyses all possible parameter set combinations and determines which is the best alternative. The final parameter set for the CB model can be found in *Tab. 5.* However, the RF model was only tuned with randomized search due to the large computational demand. The results of several runs were averaged and utilized in the final model.

| Hyperparameter | Description | Tested values | Result |
|---|---|---|---|
| **subsample** | Random selection of training data for defining splits | *range (0, 2, 0.2)* | 0.6 |
| **rsm** | Random selection of covariates defining splits | *range (0.3, 1.7, 0.2)* | 0.7 |
| **min_data_in_leaf** | Minimum data in each leaf | *range (1, 25)* | 9 |
| **learning_rate** | Used for reducing gradients step | *range (0.025, 0.2, 0.025)* | 0.125 |
| **l2_leaf_reg** | Coefficient at the L2 regularization term in the cost function | *range (0, 12)* | 0 |
| **depth** | Depth of a tree | *range (2, 20, 2)* | 8 |
| **bagging_temperature** | Defines the settings of the Bayesian bootstrap | *range (0, 1.5, 0.5)* | 0.5 |

*Table 3. Hyperparameter description, tested values and results with RandomizedSearchCV for the CatBoost model. For testing range different values were selected. The interpretation of the notation is the following: range(start value, end value, step size).*

| Hyperparameter | Description | Tested values | Result |
|---|---|---|---|
| **n_estimators** | Number of decision trees in the forest | *range (100, 2000, 20)* | 200 |
| **max_depth** | Maximum depth of the individual trees | *range (1, 100, 5)* | 16 |

| | | | |
|---|---|---|---|
| **max_features** | Number of maximum features provided to each tree | *auto, sqrt* | sqrt |
| **min_samples_split** | Minimum samples to split on an internal node | *range (1,12)* | 7 |
| **min_samples_leaf** | Minimum samples of leaf nodes | *range (1,50)* | 21 |
| **bootstrap** | Sampling with or without replacement | *true, false* | true |

*Table 4. Hyperparameter description, tested values and results with RandomizedSearchCV for the Random Forest model. For testing a range of different values were selected. The interpretation of the notation is the following: range(start value, end value, step size).*

| Hyperparameter | Tested values | Optimized value |
|---|---|---|
| **subsample** | *0.4, 0.6, 0.8* | 0.4 |
| **rsm** | *0.5, 0.7, 0.9* | 0.5 |
| **min_data_in_leaf** | *5, 9, 13* | 5 |
| **learning_rate** | *0.1, 0.125, 0.15* | 0.1 |
| **l2_leaf_reg** | *0, 4, 8* | 8 |
| **depth** | *6, 8, 10* | 6 |
| **bagging_temperature** | *0, 0.5, 1* | 0 |

*Table 5. Tested and optimized hyperparameters with GridSearchCV (CatBoost model)*

### 2.4.6 Sensitivity analysis

In general, the various input variables have varying effects on the model's target variable and prediction skill. A sensitivity analysis was performed to determine which features are the most relevant and which ones have an insignificant effect on groundwater level prediction. This approach is effective for working with big amounts of data since it explores the link between model performance and datasets, allowing for data reduction without losing information. Feature importance was estimated using a Scikit-Learn built-in tool. When producing a prediction, this approach provides scores to input characteristics, representing the relative importance of each item. This approach and built-in tool are efficient in minimizing model input data and better understanding data-prediction linkages. Implementing a sensitivity analysis and investigating the feature relevance might provide important information about the spatial patterns and trends of various input variables. This may result in data reduction, as various locations with unique meteorological, hydrological, and environmental features may have different weight factors for the input variables.

### 2.4.7 Evaluation

The models are evaluated by comparing their results to those of prior models and groundwater level observations, using different performance metrics. A sensitivity analysis was also performed to examine the significance of various input features. In order to ensure that the model is appropriately evaluated, training on full dataset must be avoided. The data was split into 50% training and 50% testing for this purpose. The model will not see the testing data in this case, which is critical to avoid overfitting to the training set. With this approach the model can be validated with approximately 2000 well locations, thereby a nation-wide groundwater map can be developed.

### 2.4.8 Evaluation metrics

The mean squared error (MSE), root mean squared error (RMSE), R-squared, and Pearsons's correlation were used to summarize the model's performance. These metrics were chosen based on past research on the use of machine learning algorithms for groundwater forecasts, as well as on machine learning model pipelines in general.

The degree of inaccuracy in statistical models is measured by the mean squared error (MSE). The average squared difference between observed and predicted values is calculated. The MSE equals zero when a model has no errors. The error is calculated by the following equation:

$$MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n} \quad (1)$$

In the equation $y_i$ denotes the i[th] observation value, $\hat{y}_i$ the corresponding prediction value and $n$ the number of observations.

Root Mean Square Error (RMSE) is the standard deviation of the prediction errors. It can be calculated by taking the square root of the MSE, as it follows:

$$RMSE = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n}} \quad (2)$$

In a regression model, R-Squared is a statistical measure of fit that shows how much variance in a dependent variable is explained by the independent variables. A score of zero indicates that the linear model is no better than the mean model, whereas a value of one indicates that the linear model is perfectly fit. The following equation can be used to compute the value:

$$R^2 = 1 - \frac{RSS}{TSS} \quad (3)$$

The sum of squares of residuals (the amount of variance not explained by the regression model) is represented by RSS, while the total sum of squares is denoted by TSS (how much variation is in the dependent variable).

The test statistic Pearson's correlation coefficient assesses the statistical link, or association, between two continuous variables. In this case the correlation is used to investigate how strong is the relationship between observed and predicted groundwater levels. The formula can be written as follows:

$$r = \frac{n(\sum y_i \hat{y}_i) - (\sum y_i)(\sum \hat{y}_i)}{\sqrt{[n\sum y_i^2 - (\sum y_i)^2][n\sum \hat{y}_i^2 - (\sum \hat{y}_i)^2]}} \quad (4)$$

The value of r ranges from -1 to 1, with 0 indicating no link between the two variables and -1 and 1 indicating a significant negative and positive relationship, respectively.

# 3. Results

## 3.1 Potential of ML models in groundwater prediction

### 3.1.1 Importance of model selection: Random Forest or CatBoost?

A generalized groundwater model was created to evaluate the influence of ML model selection. Such a model can be used to model groundwater level (and errors between the numerical model and the observations) throughout the Netherlands. This model incorporates the groundwater simulation results (*section 2.2.4*) as well as the input variables discussed in *sections 2.1.1, 2.2.2* and *2.2.3.* Sufficient groundwater level data for the total of 4002 observation wells were available. These groundwater level time series were linked to the input variables, which included soil type, land use type, elevation, distance from nearest water body, and the outputs of the existing numerical model, by combining them based on their coordinates. Meteorological variables were not acquired for each

well location but pooled into 13 separate areas and thus connected to the training data. Additionally, the same time series of hydrological observations (discharge of the Rhine and Meuse, seawater level at Haringsvliet) were used for every well location.

The characteristics of the used numerical model are described in *section 2.2.4. Fig. 5.* shows the visualization of observations and predictions for the implemented ML models and the numerical model. The plot represents the errors between the numerical simulation results and the groundwater level observations for every investigated location as well as the error of the implemented ML models, while *Tab. 6.* lists the overall assessment criteria of the model. The model predicts the observations perfectly if the values are placed on the blue line, whereas values above and below the line reflect overpredictions and underpredictions, respectively.

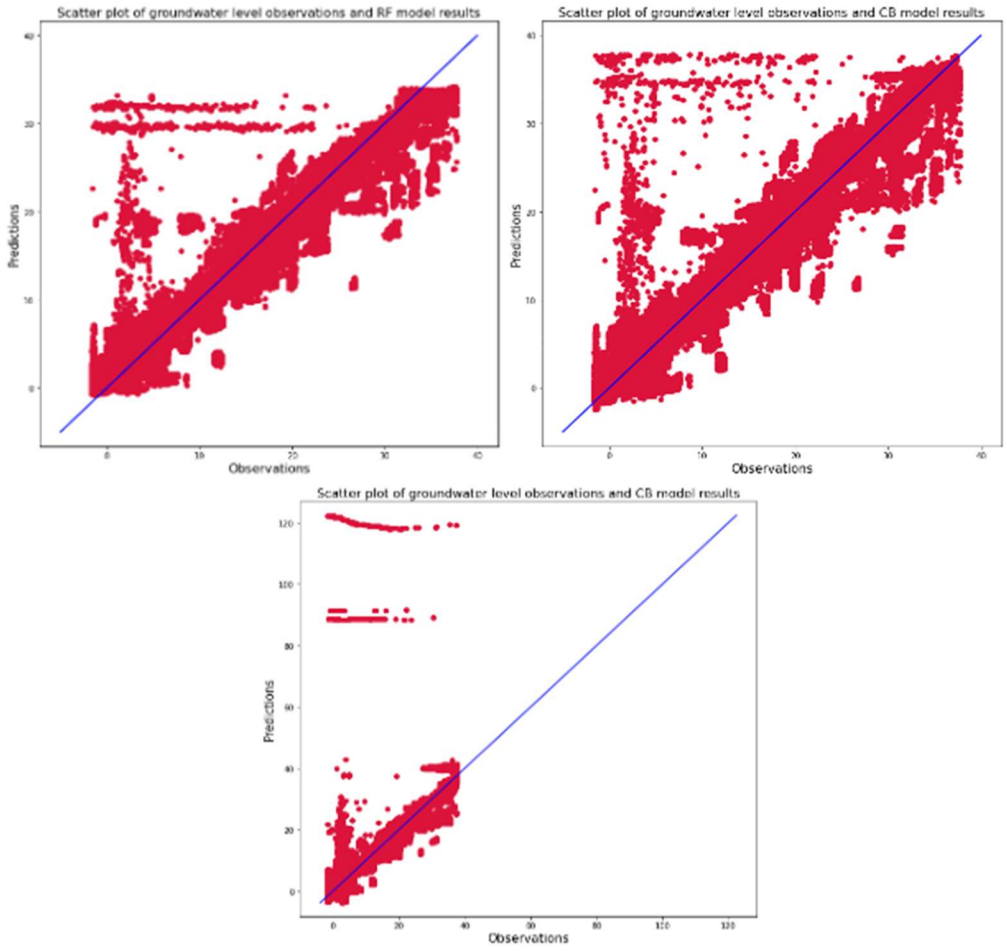
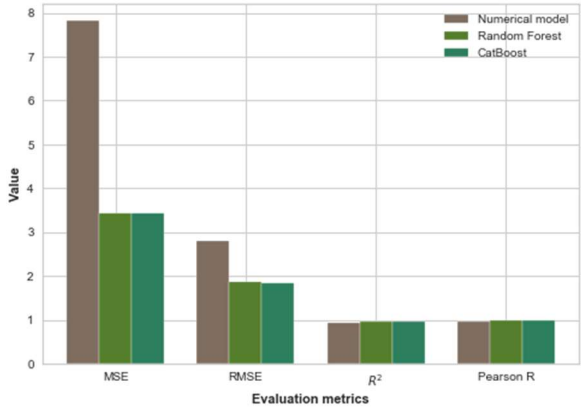



*Figure 5. Scatter plot of modelling errors. Upper left: Scatter plot of RF model. Upper right: Scatter plot of CB model. Down: Scatter plot of the used numerical model. The magnitude of errors are similar considering the two imolemented ML algorithms. The numerical model has extreme outliers, which results in the magnitude difference of the plots*

The plot (*Fig. 5.*) indicates that both ML methods can enhance the numerical model's outputs, particularly when error levels are quite high. The figure shows error values for the numerical model in the range of 120 and 90, but these values were reduced by approximately 70% and 50%, respectively. However, for very low error levels the performance of each model is relatively identical. The RF model appears to be slightly better at correcting very extreme, higher values, and the models are mostly identical at predicting low-error values. Further information can be obtained by investigating the implemented evaluation metrics listed in *section 2.4.8.* These metrics were calculated for the test

data, approximately 2000 wells, dispersed across the Netherlands. *Tab. 6.* supports that both ML algorithms could enhance the model performance in general. The MSE increases as the difference between predicted and expected values grows. The MSE value of the original numerical model ($MSE_{numerical}$ = 7.8182) was improved by roughly 55.9% using the RF model. Furthermore, the CB model managed to improve the performance of the existing numerical model and the implemented RF model, by 56.2% and 0.3% respectively. Calculating the root mean square error, which calculates average difference between the predicted values and the actual values in the dataset, is another technique to analyse how well a regression model fits. As *Fig. 6.* and *Tab. 6.* show, the RMSE value is considerably improving compared to the original value of the numerical model ($RMSE_{numerical}$ = 2.7961). The improvement is approximately 34% for both the RF and CB model, respectively.

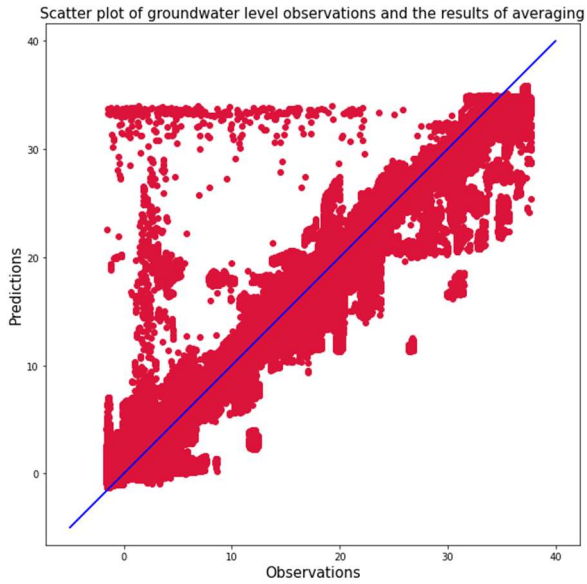



| | Numerical | RF | CB |
|---|---|---|---|
| MSE | 7.8182 | 3.445 | 3.4219 |
| RMSE | 2.7961 | 1.856 | 1.8498 |
| $R^2$ | 0.9202 | 0.9587 | 0.9627 |
| Pearson's | 0.960 | 0.983 | 0.982 |

*Table 6. Evaluation metrics of the used numerical mode and the developed ML models (RF and CB)*

*Figure 6. Barplot of the evaluation metrics for the used numerical model and the developed ML models (RF and CB)*

As a potential improvement an approach called ensemble averaging has been also utilized. Ensemble averaging is the process of constructing numerous models and integrating them to get a desired output rather than just one model in machine learning. Because the numerous errors of the models "average out", an ensemble of models frequently outperforms a single model. As *Fig. 7.* and *Tab. 7.* show that the method enhances the performance skill of both models. This is because each model is limited to learning only a portion of the structure of the data and averaging them partly compensates the errors of the others.




| | Averaged model |
|---|---|
| MSE | 3.137 |
| RMSE | 1.771 |
| $R^2$ | 0.964 |
| Pearsons's | 0.984 |

*Table 7. Evaluation metrics of the averaged model*

*Figure 7. Scatter plot of model errors: averaged ML model (including the results of RF and CB model)*

### 3.1.2 Investigating the feature importance of the general model

Another major objective of this research is to investigate how different input factors impact the model's prediction abilities and how useful they are in predicting groundwater levels. For this purpose, the feature importance of different scenarios was calculated with a built-in tool of Scikit-Learn. The implemented tool gives input features a score depending on how important they are in predicting the outcome. The higher the score, the more the specific variable is responsible for predicting the output. Using this method and examining how different characteristics impact model behaviour might assist reduce the amount of data needed for computing, thereby decreasing computational time and demand. Feature importance was calculated for both implemented ML algorithms (RF and CB). Firstly, the importance was calculated for the original setup. The calculations were done including the groundwater simulation results as well as excluding them. This was essential since the numerical model outputs already had considerable correlations with actual groundwater levels, making this characteristic much more relevant in every case than the other input variables. By removing groundwater simulations, other input variables, whose ratio is the most essential to explore, may be shown more clearly. Additionally, the data was divided into two groups to investigate the changes in feature relevance between the low and high lying parts of the Netherlands: well observations and corresponding input features when the elevation is less than 7 meters and greater than 7 meters.

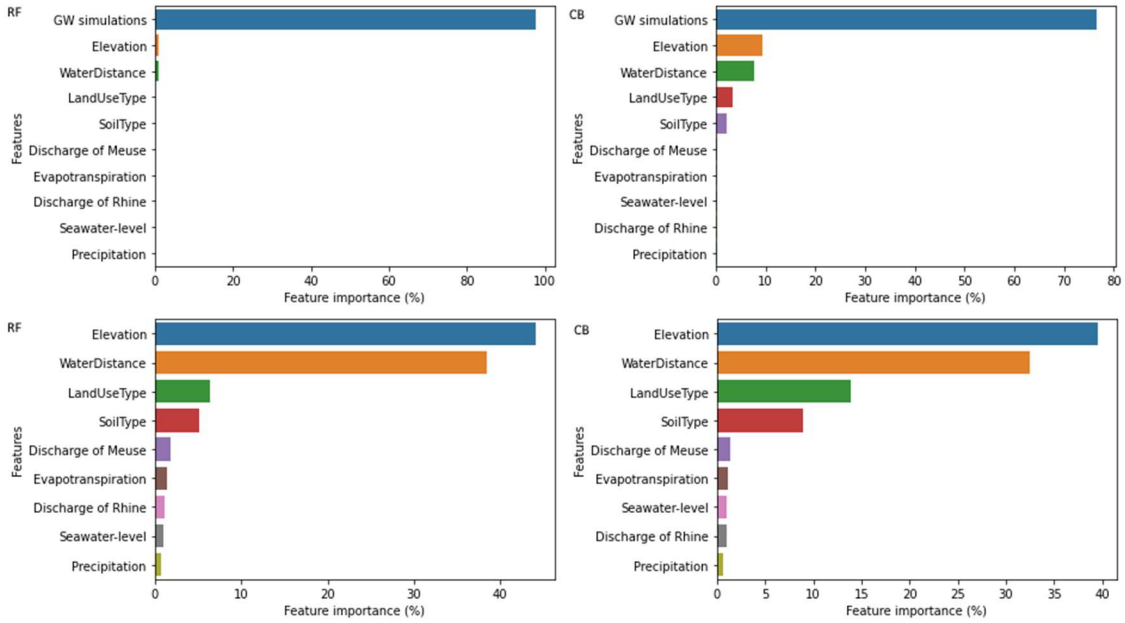



*Figure 8. Feature importance of the implemented ML models. Upper left: RF including the groundwater simulations. Upper right: CB including the groundwater simulations. Down left: RF excluding the groundwater simulations. Down right: CB excluding the groundwater simulations.*

The feature importance of the original model was calculated for both ML models. *Figure 8.* depicts the computed significance by incorporating and eliminating groundwater simulations (so the two cases are identical, the exclusion is only for visualization purposes). As the figure shows, the order of the variables is almost the same for the RF and CB models. The only difference is that the RF model considers the importance of the discharge of Rhine River slightly larger than the seawater-level, while the CB model changes these two features. The results show that for both models the most important features are the elevation and the water distance, followed by the land use type and soil type. Since these features have been used as constraints (no temporal variation) the implemented algorithms might have given them larger weight compared to the remaining variables which are changing with time. These factors are primarily responsible for defining the potential magnitude of groundwater

levels, whereas variables with temporal changes are responsible for determining smaller-scale fluctuations. The exact values of feature importance can be found in *Appendix C.*

### 3.1.3 Feature importance of low and high lying areas of the Netherlands

In addition, for both ML models, the feature significance for low and high elevated areas of the Netherlands was estimated. The chosen boundary between low and high lands was set at 7 meters above sea level, which correctly represents the country's northern and southern parts. In this case, *Fig. 9.* only shows the plots of the reduced feature importance (excluding groundwater levels) for every case. The importance of the numerical model can be found in *Appendix C.* As the graph depicts, here are no substantial changes between the two scenarios. The importance of different features is similar for both the low and high lying areas of the country. On the contrary to previous cases the CB model considers the importance of water distance slightly larger than elevation, however this might be the effect of the reduced amount of data used in this scenario. To summarize, the results demonstrate no significant changes between the two situations, indicating that the model does not account for variations in low and high-lying areas. The exact values of feature importance can be found in *Appendix C.*

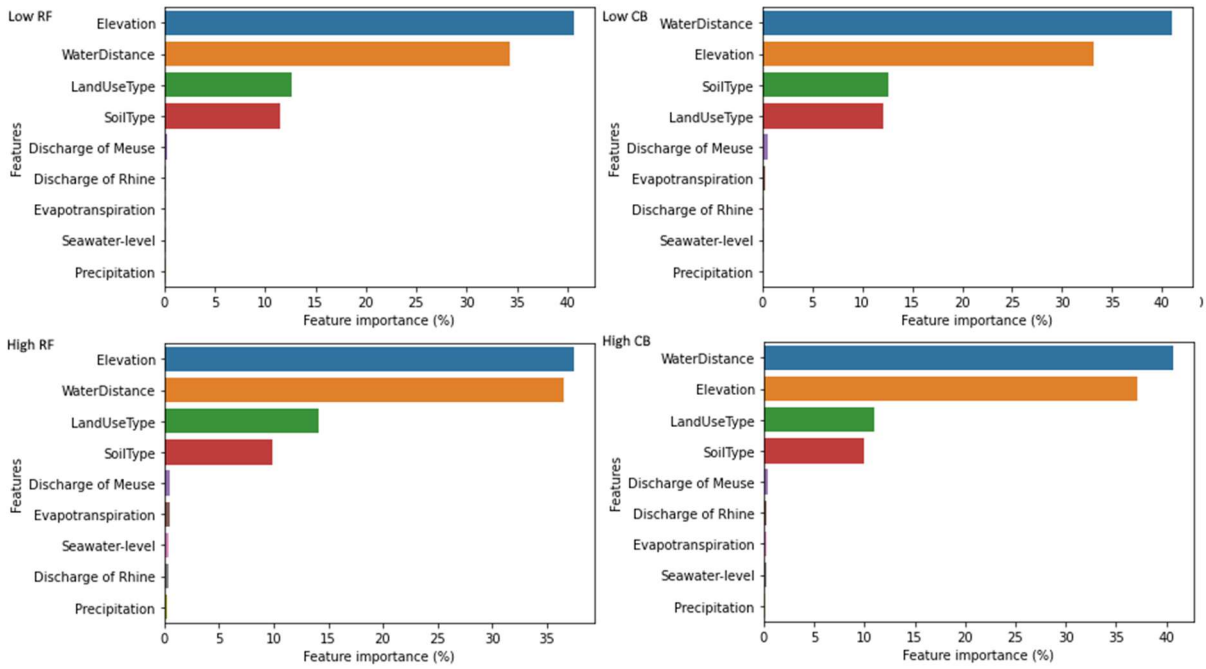



*Figure 9. Feature importance of the implemented ML models, for low and high lying areas of the Netherlands. Upper left: RF low lying area. Upper right: CB low lying area. Down left: RF high lying area. Down right: CB high lying area.*

### 3.1.4 Averaged model to investigate the possible impact of timeseries

Furthermore, a simplified model was created to study the impacts of timeseries and to provide a basic overview of the ML models' capability to estimate groundwater levels. The available input data was averaged for each observation well site throughout the whole examined time period for this model (in most of the cases approximately from 1980 to 2019, however at some locations data was not available for the entire time frame). Similar to the general model presented in *section 3.1.1* the 5% lowest and highest groundwater level measurements were eliminated to better represent the majority of the data, resulting in 3620 wells and the corresponding averaged data being utilized for this model. During the training process 70% of the data was used for training and 30% for testing. Since the results were satisfactory without a specific hyperparameter tuning for this scenario, no separate tuning was carried out.

*Tab. 8.* and *Fig. 11.* shows the calculated evaluation metrics for this scenario. The results are showing that both ML algorithms are capable of significantly improving the predictions of the numerical model. As *Fig. 10.* shows that the utilized ML methods correct severe error values while also preserving low error values near to the error margin of zero. Thereby, considering the evaluation metrics and the error scatter plots, it can be safely concluded that both ML algorithms are doing great in predicting averaged groundwater levels. Since these findings are based on averaging several decades of observations into a single variable, this model provides a very broad picture of the possibilities of ML approaches in such contexts. The results and this simplified model, on the other hand, can be used as a starting point for groundwater level modelling in unknown locations and as a base for future water management planning.

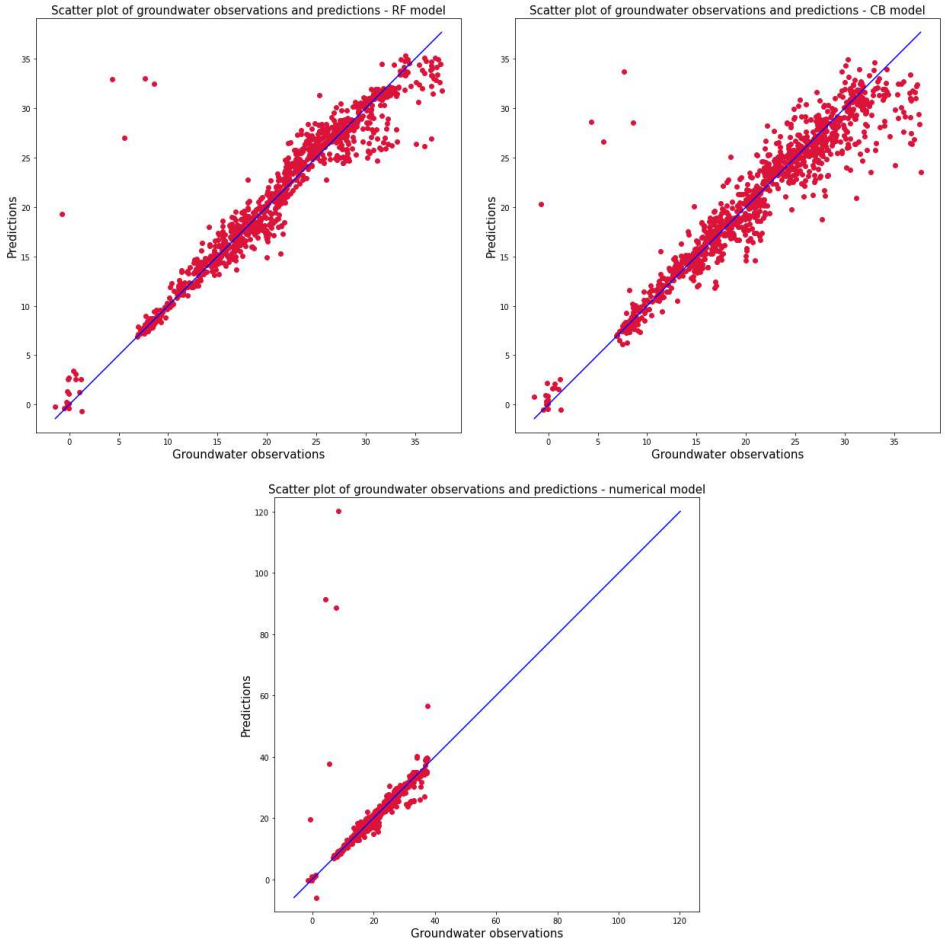
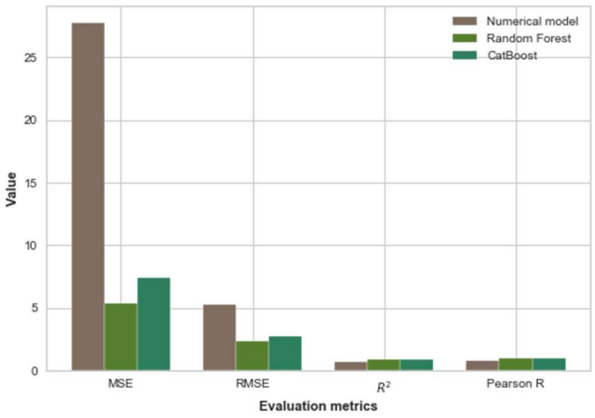



*Figure 10. Scatter plot of modelling errors – simplified, averaged scenario. Upper left: Scatter plot of RF model. Upper right: Scatter plot of CB model. Down: Scatter plot of the used numerical model.*

Additionally, the feature importance has been calculated for both ML models. This was essential in order to compare the simplified model's importance to that of the original, generalized cases. Due to the absence of temporal characteristics in this model (timeseries were averaged), novel combinations and relationships of feature importance might emerge. The reason for this is that there are input variables that do not change over time, such as land use type, soil type, elevation, and distance from the nearest water body. The algorithms learn that these variables are always the same for a well location, thereby assigning them a higher importance than the actual value should be. This limitation of the general model is further discussed in *section 4.1* and *4.5.* Groundwater level simulations were also included as input data throughout the modelling process, yet due to their apparent and considerable relevance and to better illustrate the ratio of other characteristics, they

were removed from the plot (*Fig. 12.*). The importance of groundwater level simulations was 97.497% and 72.656% for RF and CB models, respectively. As *Fig. 12.* shows the feature importance excluding the timeseries are significantly different compared to the case discussed in *section 3.1.2* and *3.1.3.* The general model did not consider the temporally changing variables that significant as the simplified model. Precipitation (for both algorithms) became the most important variable followed by the seawater-level and evapotranspiration. This shows a realistic picture of actual hydrological processes since the precipitation is one of the most important influential factors of groundwater level dynamics. The importance of soil type, land use type and elevation are considerably smaller compared to the general model, although water distance is still one of the most important parameters. The importance and relevance of water distance is obvious but incorporating it into such models is not straightforward. Water distance as an input feature might be an effective way to improve model performance and facilitate modelling in general in unknown places in future research




| | RF | CB | Numerical |
|---|---|---|---|
| MSE | 5.372 | 7.394 | 27.747 |
| RMSE | 2.318 | 2.72 | 5.268 |
| $R^2$ | 0.905 | 0.86 | 0.64 |
| Pearson's | 0.956 | 0.938 | 0.805 |

*Table 8. Evaluation metrics of the simplified model*

*Figure 11. Barplot of the evaluation metrics for the used numerical model and the simplified ML models (RF and CB)*

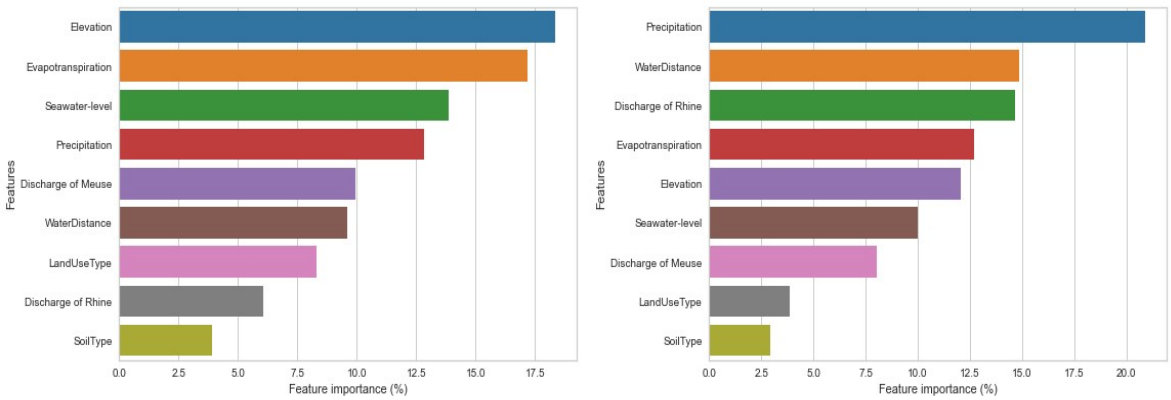



*Figure 12. Feature importance plot of the simplified ML models (left: RF, right: CB)*

Furthermore, similarly to the generalized model the feature importance has been calculated for low and high lying areas of the Netherlands as well. *Fig. 13.* shows the relevance values of these calculations. For low lying areas the elevation, evapotranspiration, sea-level and precipitation are the most significant features. In these areas the groundwater level values are mostly lower compared to the high lying areas, thereby small elevation differences can cause bigger differences in groundwater levels by magnitude. In addition, these areas are located in the northern part of the country, hence the larger importance of the sea-level and lower importance of river discharges. Contrary, the importance of sea-level in high-lying areas are significantly lower, while the importance of river discharge (especially true for the Rhine) is larger. The recharge of groundwater is mostly dependent

on precipitation in higher areas, which is in line with the realistic explanation and the expected outcome of the model.

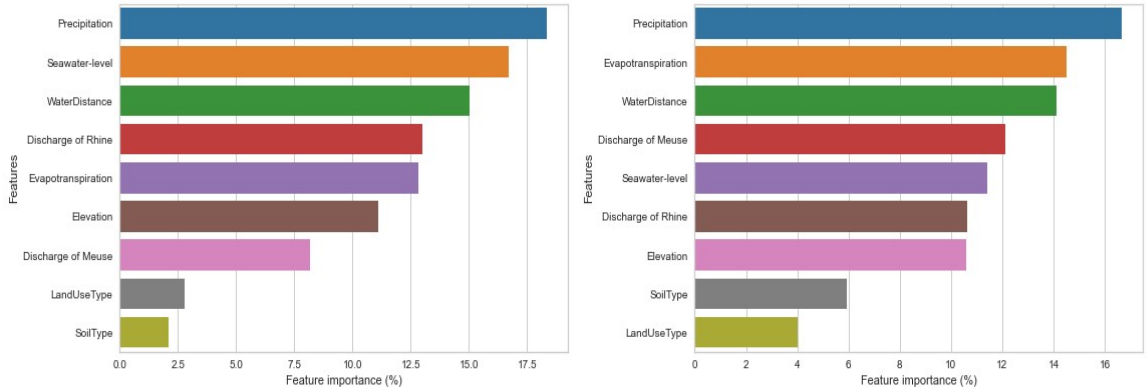



*Figure 13. Feature importance for the simplified ML models (left: RF model – low areas, right: RF model – high areas)*

## 3.2 Spatial variability of prediction error

To be able to build a realistic generalized groundwater level model and to check the limitations, a thorough spatial analysis is required. With such an analysis the spatial variability of the error (in this case mainly the MSE) and other evaluation metrics can be assessed regionally. It is necessary to determine whether the inaccuracies have a general spatial trend. By answering this question different regions can be separated, and the performance capability of the model can be determined for distinct geographical locations.

### 3.2.2 General trend in prediction error

For this reason, to better represent the results spatially, *Fig. 14.* shows the spatial density of errors between ML models and real observation values. The colour ramp represents the number of values in a pixel. Both plots appear to be very similar. The majority of the errors are clustered at the zero margin, and there is no valuable information to be gained from the scattered outliers. Thereby, a residual plot was created to provide a better understanding and to see if there were any common patterns in the model errors. The CB model's errors were subtracted from the RF model's errors and displayed against actual groundwater level measurements. This means that a negative number indicates a bigger CB model error, whereas a positive value implies a larger RF model error. As *Fig 15*. shows, and a slope value of m = 0.43 indicates, generally the errors of the RF model are higher compared to the CB model. However, both models have very similar errors when lower groundwater values are considered.

To investigate the occurrence of a general trend in the performance of the model spatially, the average MSE was calculated for every different well location (the total of 1837 wells). Calculating the average was necessary because the representation of spatial and temporal variability together is out of the scope of this research project. *Fig. 16.* shows the various MSE values for each well site, as well as the related heatmap, to better demonstrate spatial variability. By investigating the figure, it's noticeable that both models have a fairly comparable geographic distribution of error. However, the heatmap shows that both models are working with very limited errors in the northern regions of the country (i.e., lower groundwater levels), and generally larger error rates in the southern provinces (i.e., larger groundwater levels on average). The findings show that the constructed model is good at predicting groundwater levels for values that occur more frequently, but not so effective at modelling more uncommon and extreme values.

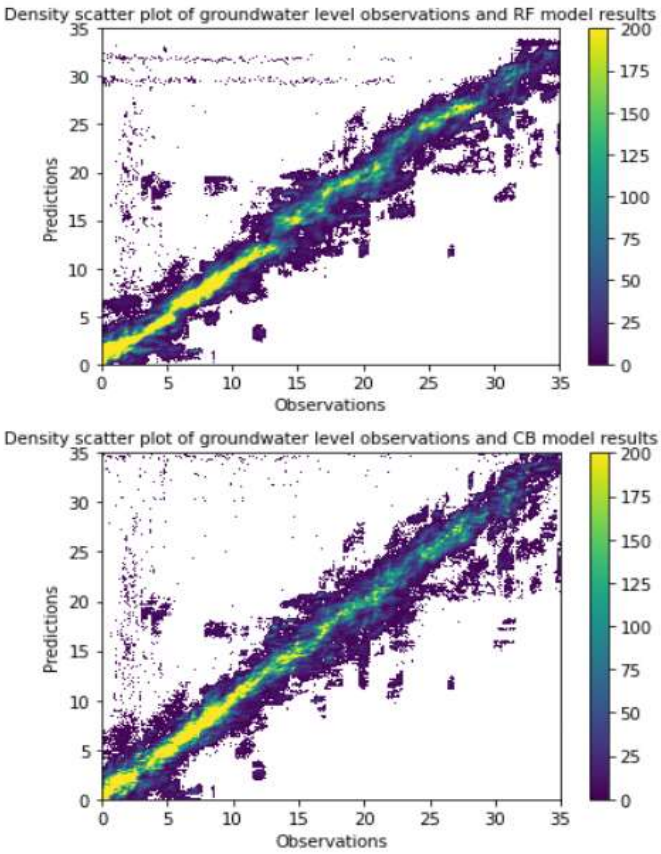



*Figure 14. Density scatter plots of the groundwater observations and implemented ML algorithms (RF and CB, respectively)*

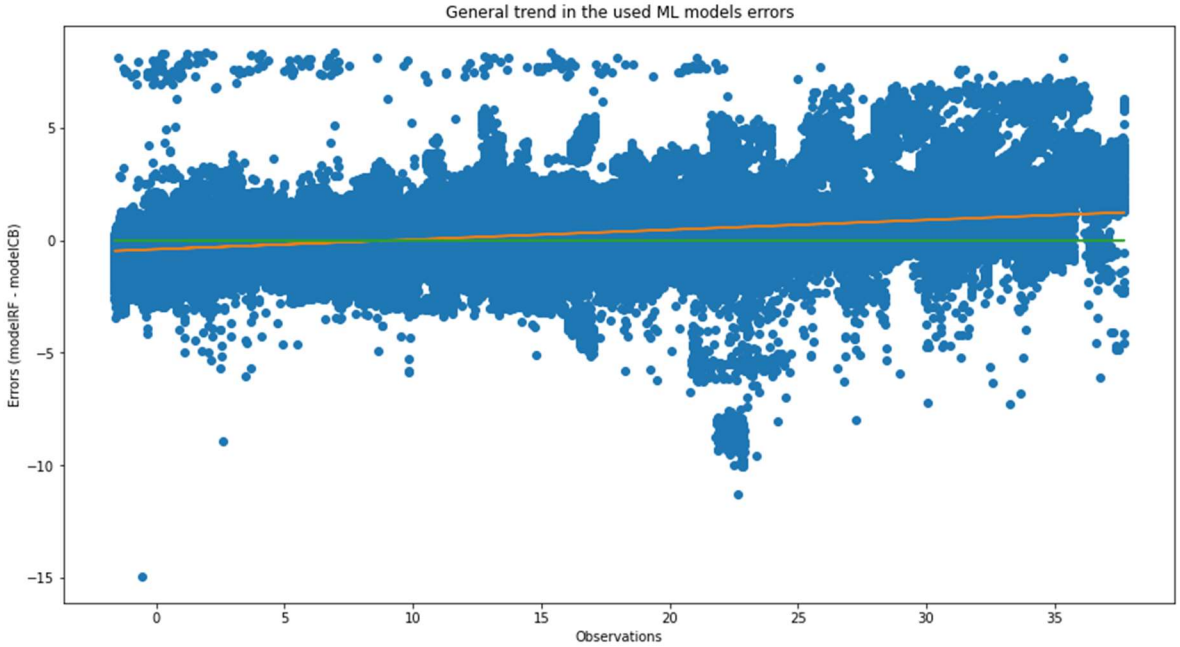



*Figure 15. Plot of residuals for the errors between RF and CB models*

Additionally, 8 distinct well locations (*Fig. 17.*) were selected randomly with the single condition of being geographically separated to provide a thorough picture of different models' spatial behaviour and to give a general overview about model performance at random locations. *Fig. 18.* and

*19.* shows the time series of the 4 of these previously selected sites (the remaining can be found in *Appendix D.*). The time series are including five separate variables, which are the following: actual groundwater observations, simulation results of the numerical model, results of the RF model, results of the CB model and the results of the averaged model obtained from RF and CB models. Considering the time series, it can be ascertained, that all the ML models are more sensitive to weekly fluctuations and thereby better represent the actual groundwater level dynamics. In most of the cases the results are properly representing the temporal variability of groundwater levels, however, as a disadvantage of the utilized ML models, the time series are shifted on some occasions. These characteristics can be the effect of two distinct attributes: 1) The results of the numerical model are already shifted in some degree due to modelling errors, and since this is the most important feature for both ML models the results will be close to these values. 2) During the modelling several features were used which were either categorical (i.e., soil type or land use type) or fixed values (i.e., elevation or distance from closest water body). Since these variables are not changing with time the weight of these values is increased.

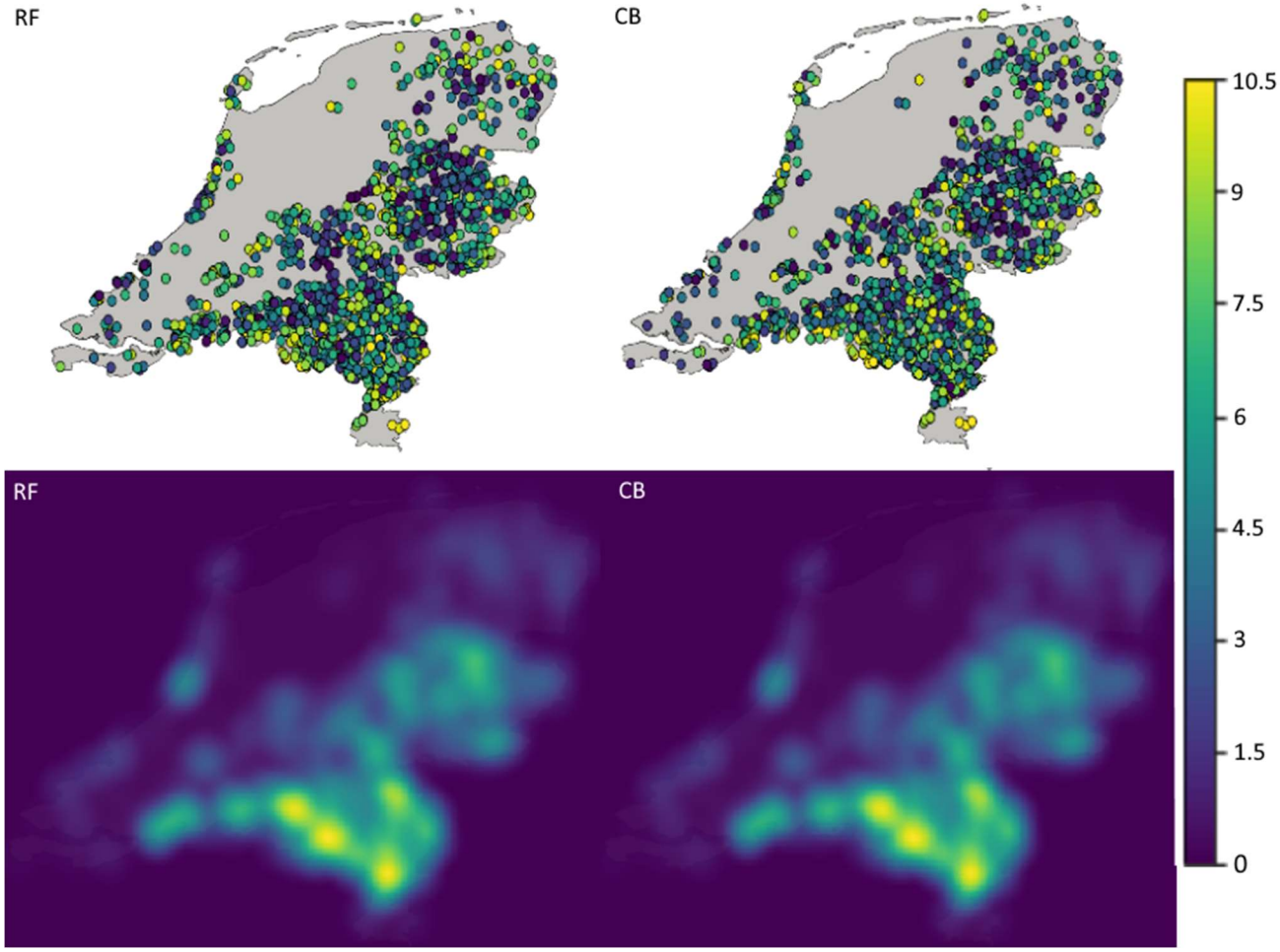



*Figure 16. Upper left: Point map of wells representing different MSE values for RF model. Upper right: Point map of wells representing different MSE values for CB model. Down left: Heatmap of RF model MSE. Down right: Heatmap of CB model MSE.*

As a result, if the model learns during the training process that one or more of the constrained variables belong to a specific groundwater level value, predictions will be erroneous in cases where the groundwater value is significantly different, but the categorical or fixed variables are similar to the previous case. The established modelling resolution of 250 meters, which might be a reason for such errors, could be one explanation for this inaccuracy *Fig. 18.* depicts an example for this phenomenon:

the MSE of the existing numerical model is lower compared to all the ML models, however the Peason's correlation is significantly higher for the ML algorithms. Evaluation metrics, including the mean squared error and Pearson's correlation can be found in *Appendix E.* This limitation is further discussed in *section 4.5.*

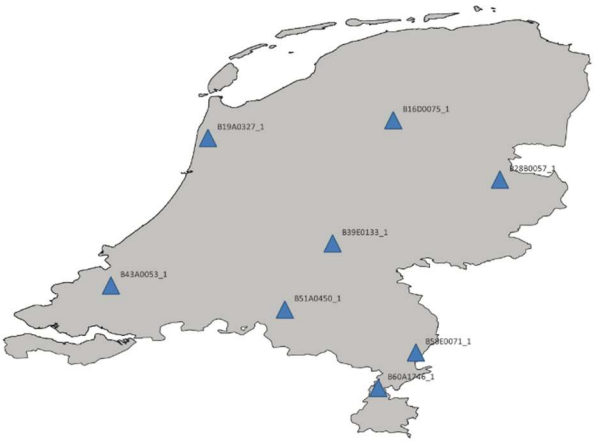
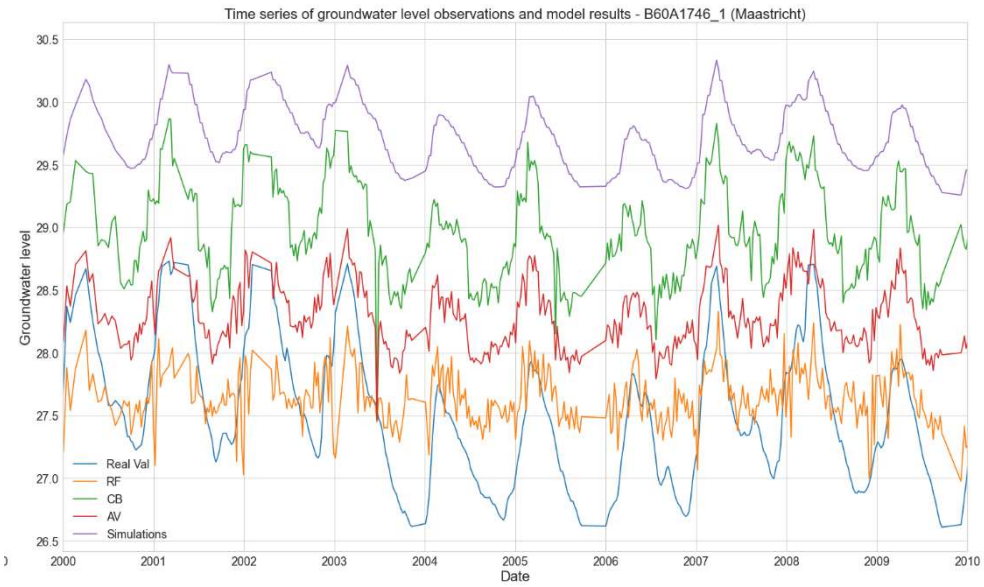



*Figure 17. Randomly selected well locations throughout the Netherlands.*



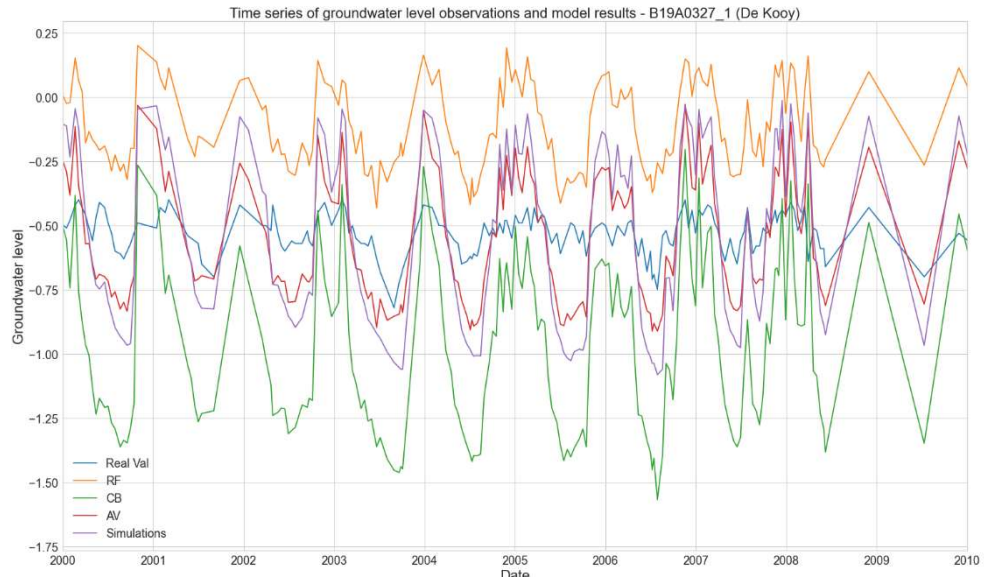




*Figure 18. Time series of 2 different well locations. Up: Maastricht region (well ID: B60A1746_1); Down: De Kooy region (well ID: B19A0327_1).*

As seen in *Fig. 18.* and *19.*, both ML models can make highly accurate predictions, but they can also be wrong by a significant margin. One possible reason for this is the build-up of these algorithms: the fundamental distinction between random forests and CatBoost (an in gradient boosting algorithms in general) is in the way decision trees are built and aggregated. CB models are building decision trees additively, while an RF model combines decision trees together to give the output. Based on the figures, it can be concluded that ML approaches are better at detecting small-scale oscillations in groundwater levels compared to the numerical model, owing to the inclusion of hydrological and meteorological factors. However, due to the large amount of data erroneous values cannot be completely avoided. Significant errors in the groundwater simulation dataset, as well as missing or false measurements from the target (groundwater level observations) and input (meteorological and hydrological observations) data as well as the implemented resolution (250m), can result in inaccurate predictions that are not directly dependent on location. Further data correction and the use of a higher resolution might correct or at least mitigate these problems.

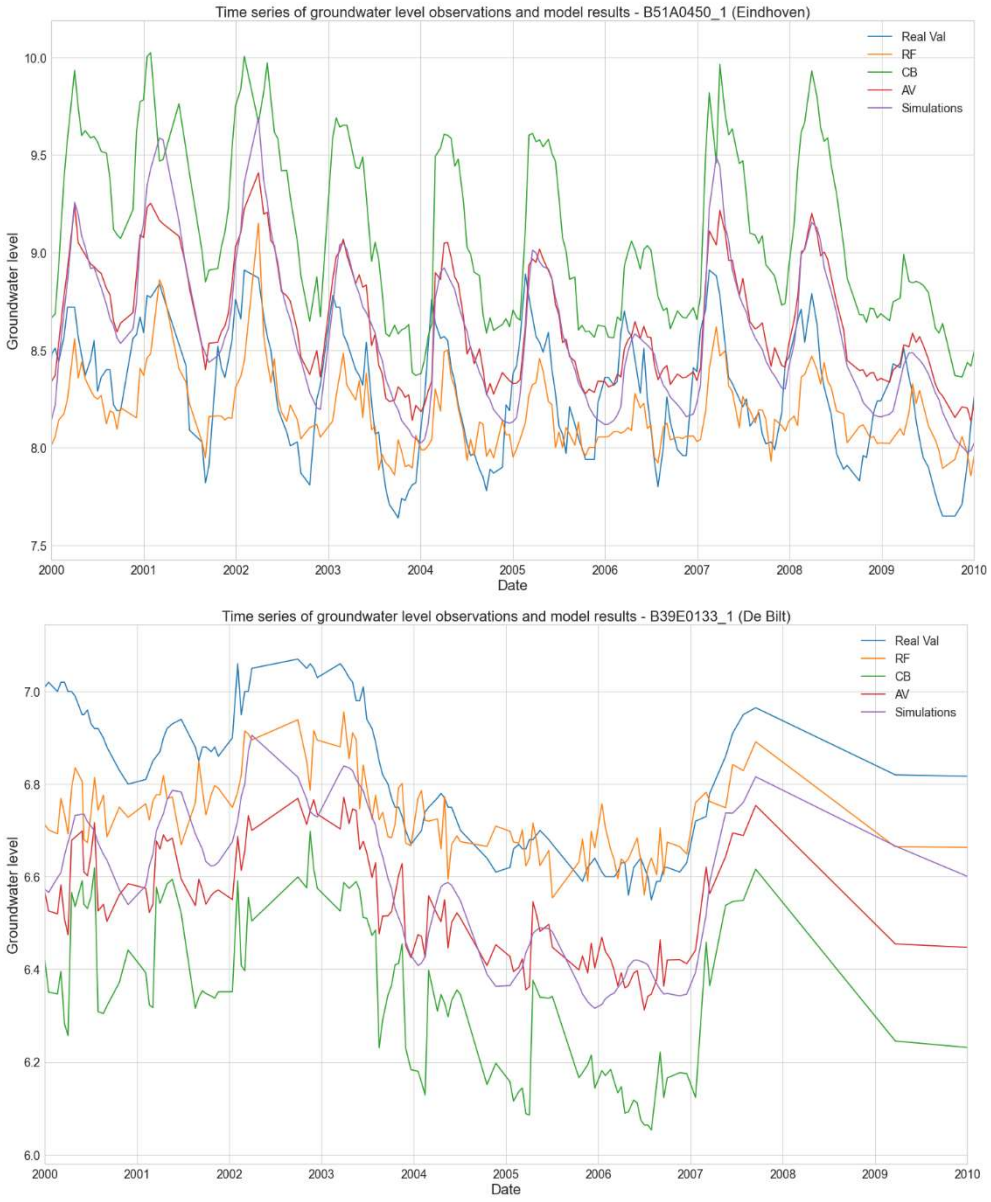



*Figure 19. Time series of 2 different well locations. Up: Eindhoven region (well ID: B51A0450_1); Down: De Bilt region (well ID: B39E0133_1)*

## 3.3 Error case

### 3.3.1 General model

In addition to the previous model, an error model was created to see if the available data and, more generally, machine learning techniques could be used to estimate the error between the numerical model and the groundwater observations. This approach can be valuable because, after the model has been trained (essentially after calculating the errors), it can be used to predict errors and hence determine groundwater levels without the need for actual measurements. For this purpose, the error between the numerical model and the actual groundwater measurements has been computed in the following way:

$$Errors = Groundwater\ simulation\ results - Groundwater\ observations$$

In the first scenario, this error was estimated for all the sites for which data was available. (i.e., all the well locations, approximately 4000, for all the inspected time period). Similarly, to the previous model the extreme groundwater level values (5% tails of the dataset) were removed to improve the model's performance. The available data was split into 50% training and 50% testing. The used input variables were the same as in the previous cases, thereby all the meteorological, hydrological, and environmental variables were implemented as well as the numerical model's results. Hyperparameters have been tuned to get the best possible results, although only the randomized search was implemented for both ML models.

| Metrics | RF model | CB model |
|---|---|---|
| MSE | 7.625 | 8.122 |
| RMSE | 2.761 | 2.845 |
| $R^2$ | -27.423 | -10.451 |
| Pearson's correlation | 0.158 | 0.087 |

*Table 9. Evaluation metrics of the error model*

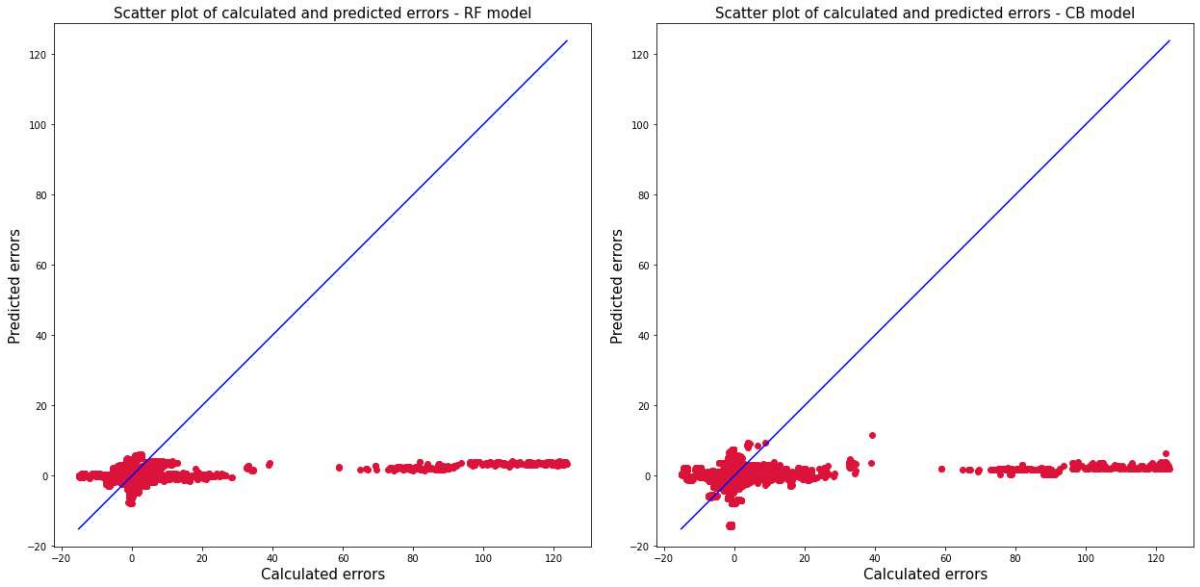



*Figure 20. Scatter plot of calculated and predicted errors for both utilized ML models (RF and CB)*

As *Tab. 9.* and *Fig. 20.* represents, the performance of the model with the aforementioned setting is insufficient. The model is not capable to learn relationships and connections in the data, thereby the predictions are constantly close to zero. Since the majority of the calculated errors are very close to zero, the model becomes imbalanced and not sensitive to larger, more extreme, and

infrequent groundwater level values. In addition, considering the plots in *Fig. 21.* It can be observed that the residuals (ratio between calculated and the difference between calculated and predicted errors) are not randomly distributed around the zero line in both cases. This generic pattern demonstrates that the model is unable to extract sufficient information from the implemented input features, making it unsuitable for error modelling in the current configuration.

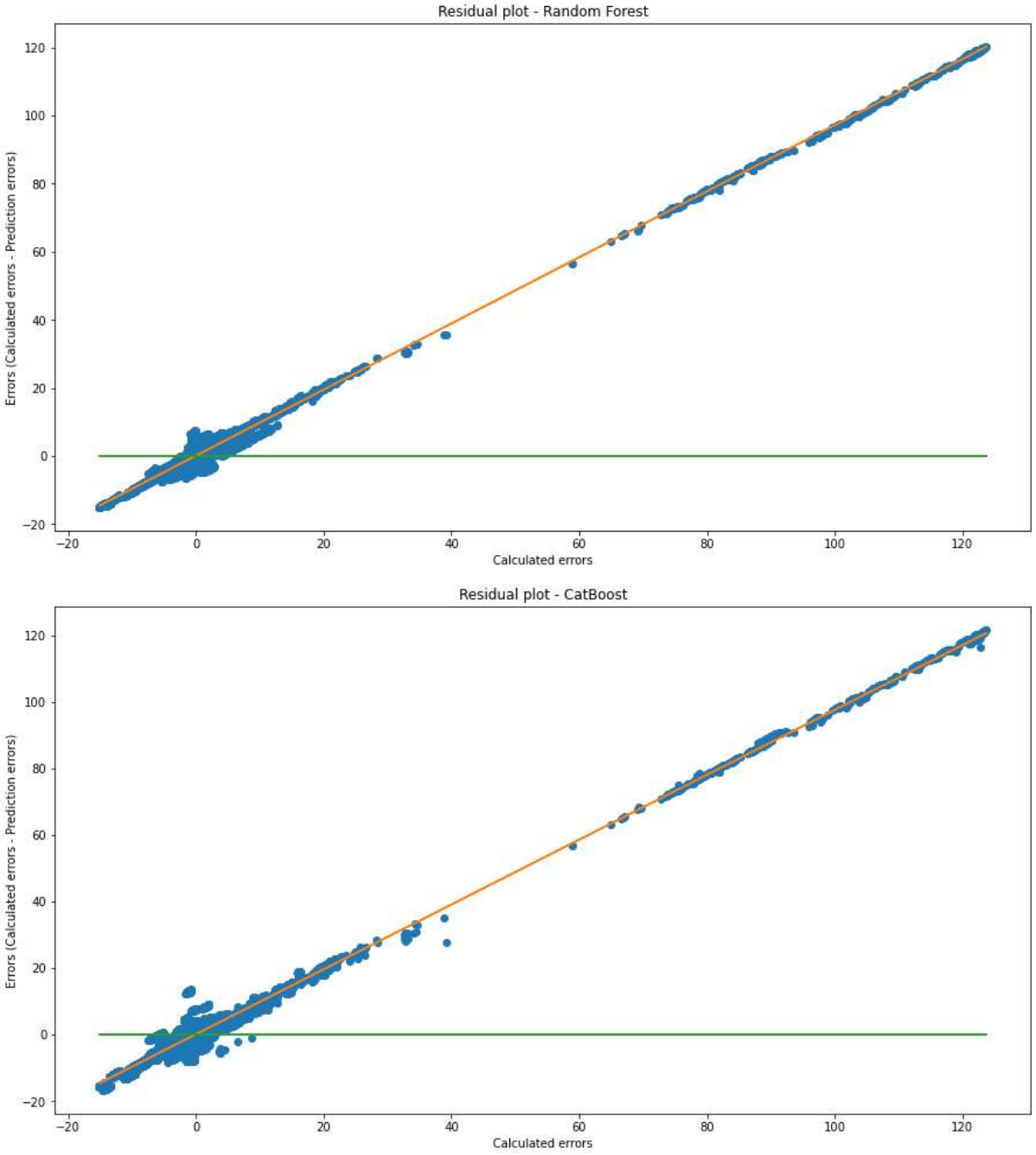



*Figure 21. Residual plots for the implemented ML methods*

### 3.3.2 Simplified model

A simplified, averaged model was created similarly to the previous cases in order to overcome this and produce a generalized ML model capable of estimating the errors between groundwater level simulations and measurements. This model was averaged throughout the investigated time period to provide a single set of general input feature values for each unique well site. This scenario also included the deletion of the 5% tails of severe groundwater level measurement readings. In order to

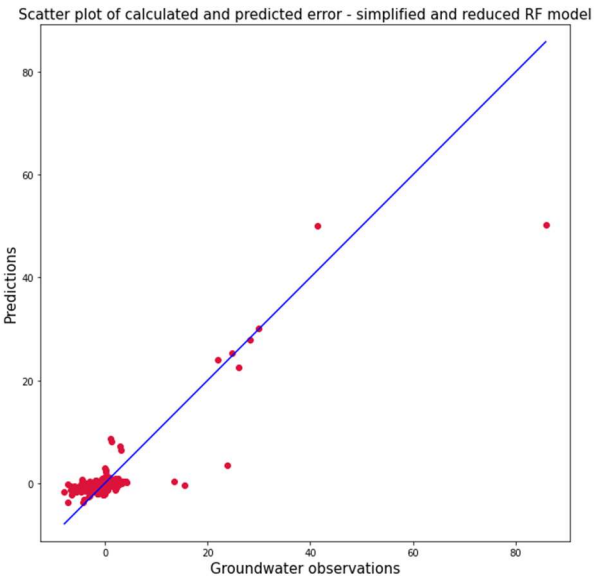
improve the performance, the data was randomly split into 70% training and 30% testing (in general more training data improves the prediction skill of such models). The model performance is similar compared to the previous scenario, where all available data and timeseries were included without averaging, as shown in *Tab. 10.* and the scatter and residual plots in *Appendix F*. The MSE and RMSE measures are worse, but the correlation has improved significantly, which might be due to the severely reduced data. The performance of the RF model is considerably better compared to the CB model, hence for the upcoming scenarios only the RF model's results are presented.

| Metrics | RF model | CB model |
|---|---|---|
| MSE | 8.111 | 14.202 |
| RMSE | 2.848 | 3.769 |
| $R^2$ | -2.441 | -13.342 |
| Pearson's correlation | 0.857 | 0.398 |

*Table 10. Evaluation metrics of the simplified and averaged error model*

### 3.3.3 Simplified and reduced model

According to the prior findings, a general trend in the predicted error between groundwater level calculations and measurements was detected. For a sufficiently working model this pattern has to be eliminated or at least lowered. Potentially, one of the reasons what could cause this trend is that the groundwater level data is still deficient for more extreme values. As a result, following the already implemented approach 15% of the tails (so in total 30% of the data) was eliminated. The train-test data ratio, as well as the parameters and input variables, remained unchanged. *Tab. 11.* and *Fig. 22.* shows the evaluation metrics and hence the model in general is improved by approximately 50% considering the MSE value. The outcome of this scenario demonstrates that it may be feasible to predict the errors between simulation and actual data, but only under particular conditions. The model is unbalanced due to a lack of data for high groundwater level values and removing these might enhance prediction skills. However, the residual plots for this case in *Appendix F.* proves, that there's still a pattern existing, thus the model prediction errors are still not completely random.




| Metrics | RF model |
|---|---|
| MSE | 4.006 |
| RMSE | 2.001 |
| $R^2$ | 0.611 |
| Pearson's correlation | 0.883 |

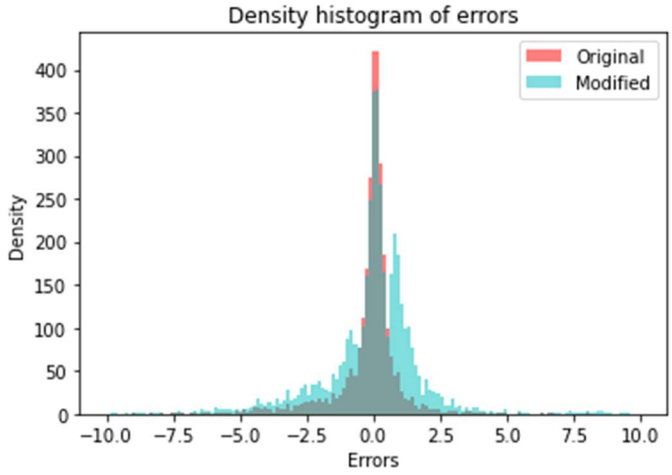
*Table 11. Evaluation metrics of the simplified, averaged, and corrected error model*

*Figure 22. Scatter plot of calculated and predicted errors – simplified and reduced RF model*

### 3.3.4 Simplified, reduced model with the implementation of SMOGN approach

Furthermore, in order to improve the model and eliminate or lower the discovered trend in prediction error values, a technique called SMOGN was used. Skewed distributions with a long tail are

common in real-world datasets. This approach is useful for prediction problems when regression is applicable, but the values to be predicted are infrequent or unusual. This can also be a good alternative to log converting a skewed response variable, which in this case due to the occurrence of negative values is not possible (Branco et al., 2017). As a result, the data values that occurred in the minority (i.e., large error values related to deep groundwater levels) are oversampled in order to better represent a normal distribution. *Fig. 23.* shows the original and the modified density distribution of the implemented variables. *Tab. 12.* and *Fig. 24* depicts the evaluation metrics and the error scatter plot of this approach, respectively. The results of the evaluation metrics are considerably worse compared to the previous, reduced scenario. The approach might be helpful correcting some errors in the minority class, but on the other hand in worsens the prediction skill in the majority class, thus the skills of the developed model in general.




| Metrics | RF model |
|---|---|
| MSE | 11.612 |
| RMSE | 3.408 |
| $R^2$ | 0.497 |
| Pearson's correlation | 0.831 |

*Table 12. Evaluation metrics of the simplified, averaged, and corrected error model with implementing the SMOGN technique*

*Figure 23. Density histogram of model of the errors used as target variables for both cases*

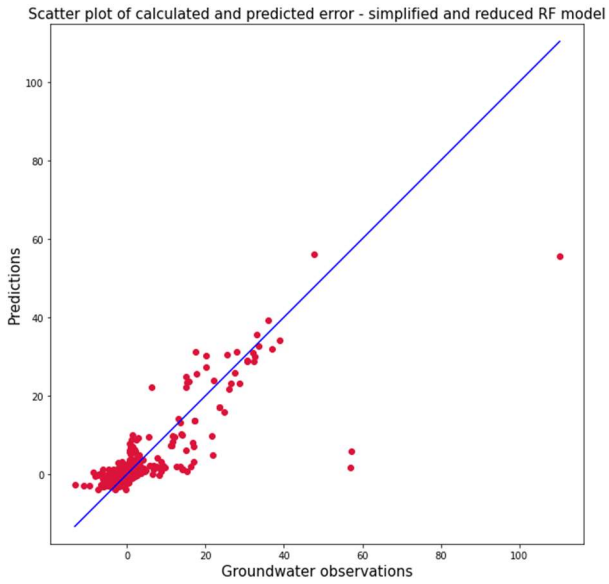



*Figure 24. Scatter plot of calculated and predicted errors – SMOGN*

### 3.3.5 Feature importance of the error model

Similarly, to the previous cases the feature importance was also calculated, to examine which input variables might have the most crucial impact on the outcomes and the modelling performance in general. To properly demonstrate these values, the best performing model have been chosen. Groundwater level simulations were removed in order to better illustrate the findings, as they had a

substantially greater feature relevance value than the remaining input variables (i.e., 89.47%). The distribution of the remaining, approximately 10.53% of the feature importance can be found in *Fig. 25.* In lights of this figure, it can be concluded that the feature importance distribution is similar considering the previous cases for groundwater level predictions. In general, meteorological factors and physical features are playing a more important role in such models.

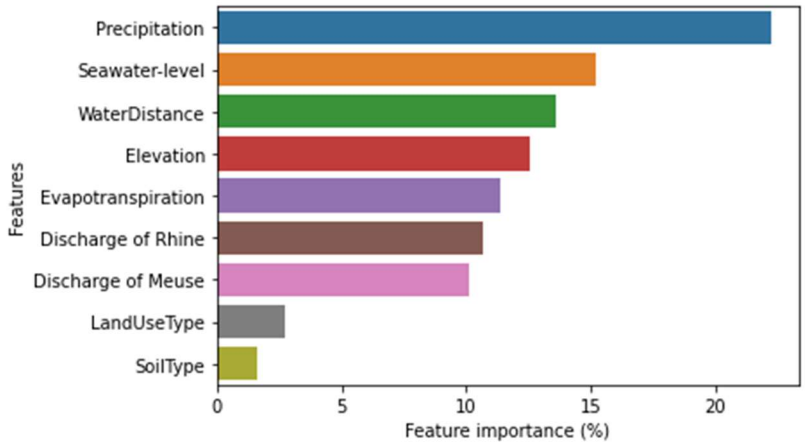



*Figure 25. Feature importance values of error model (for the case discussed in section 3.3.3)*

## 3.4 Groundwater map of the Netherlands

A groundwater map was also constructed to provide a better visual depiction of the research's primary results. Two maps were created with a 250m resolution using the outputs of both ML algorithms. The primary model's results, discussed in *section 3.1*, were employed. For the time period under consideration, the values were averaged. Then, for each individual site, the data from the summer months (June, July, and August) were selected and averaged. As a result, the map depicts groundwater levels in the Netherlands overall during normal summer circumstances. Typical summer conditions were chosen based on the importance of groundwater depth throughout the year. During the summer, the nation may experience periods of insufficient rainfall, resulting in severe droughts. This might cause major issues with irrigation and agricultural productivity in general, therefore knowing the typical groundwater levels in such settings could be useful as a starting point for future modelling and mitigation purposes. As *Fig. 26.* and *Fig. 27.* shows, that if the groundwater levels are visualized in a continuous scale there are no significant differences between the two utilized ML models. The figures accurately depict the anticipated outcomes: shallow groundwater levels in the north-western and northern parts of the country, and deeper levels in the south and east. The results also representing the locations of main river bodies (i.e., Rhine and Meuse Rivers) and the deeper levels around the Hoge Veluwe National Park in Gelderland in the middle of the country. The CB model is more sensitive to groundwater extremes, since the predictions are scattered on a larger scale compared to the RF model. To better represent the differences between the two implemented models the groundwater levels has been divided by quantiles as well. Groundwater levels were also split by quantiles to better show the differences between the two deployed models. This method helps in the better visualization of lower groundwater levels in the country's northern regions, allowing to explore smaller scale variations between groundwater levels and the two models. *Fig. 28.* and *Fig. 29.* depicts the groundwater maps, scaled by quantiles for RF and CB model, respectively. For the deeper groundwater levels in the southern part, both models provide quite comparable estimates, therefore there is no noticeable difference in their performance. From *Fig. 28.* it can be concluded that the RF model is not completely accurate in some parts in the northern regions and overpredicting the anticipated groundwater levels. This is mostly true for the region close to the city of Almere and the area between Amsterdam and Rotterdam. The majority of these areas are lower than the actual sea level, however the model predicts higher levels for some parts. On the other hand, the CB model has a significantly better performance regarding these areas.

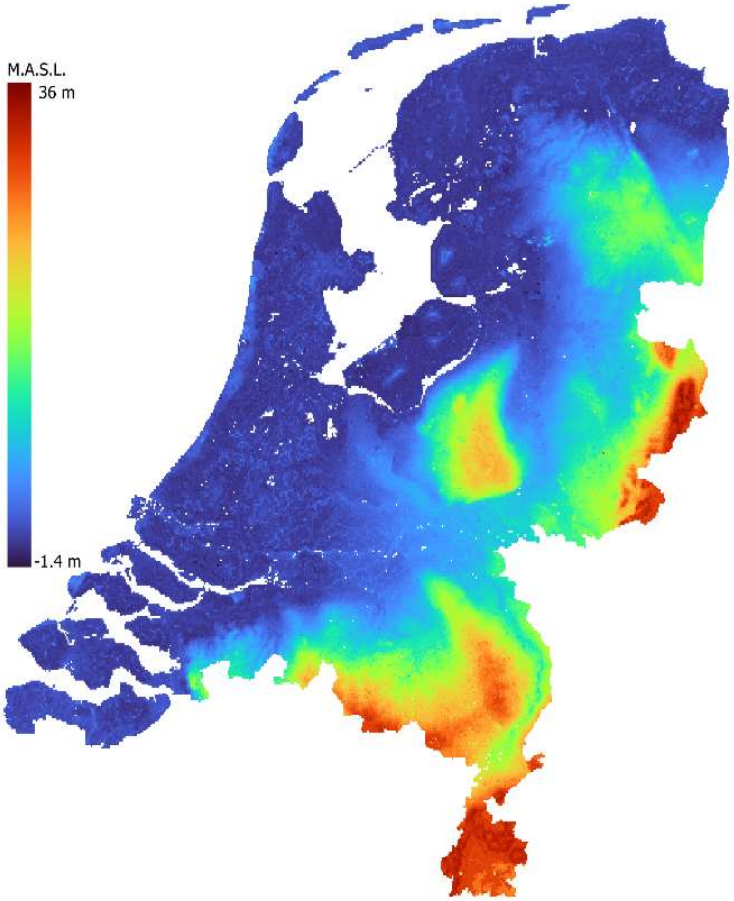



*Figure 26. Groundwater levels predicted by RF model for the whole area of the Netherlands. This continous map represents the main characteristics of grounwater level depth in the country*

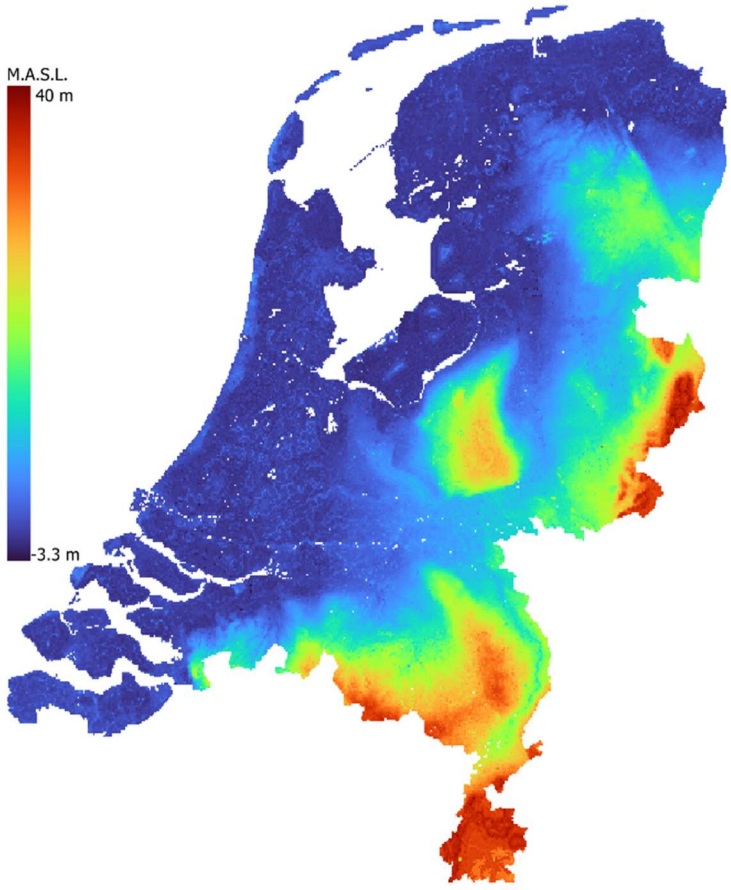



*Figure 27. Groundwater levels predicted by CB model for the whole area of the Netherlands. This continous map represents the main characteristics of grounwater level depth in the country*

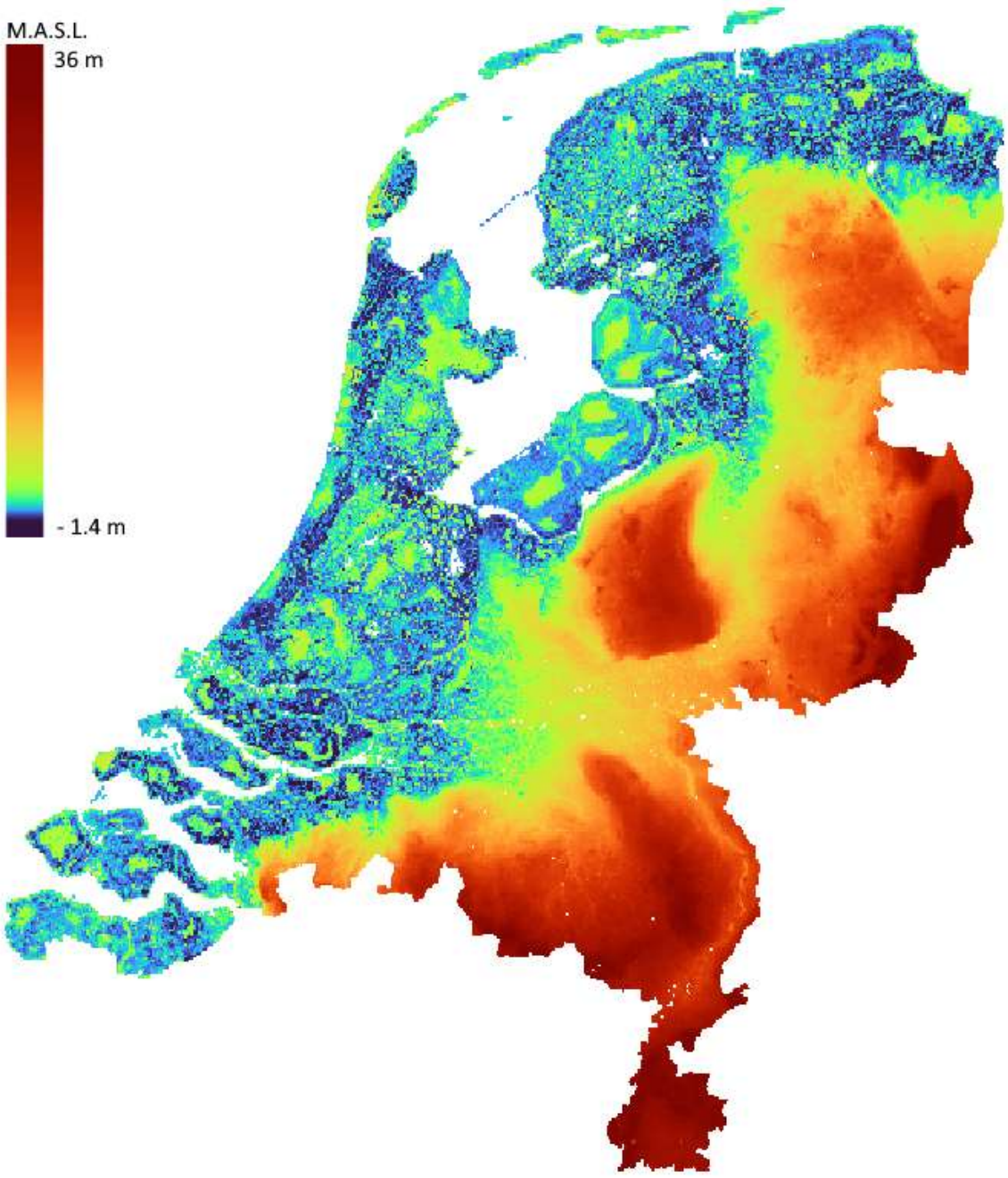



*Figure 28. Groundwater levels predicted by RF model for the whole area of the Netherlands. This quantile map represents the smaller scale variations of grounwater level depth in the country, especially in the northern parts with less deep groundwater levels*

Another advantage of the CB model is that it's more sensitive to represent the actual (deeper) groundwater levels around the dunes close to the North Sea as well as to the higher groundwater levels for some parts of the Frisian islands. In summary, the maps are a good representation of the general conditions of groundwater level distribution of the Netherlands. The implemented ML algorithms are sensitive to regional differences, as well as more local, smaller scale variances. The implemented methods and results are giving a fairly accurate representation of the actual groundwater level conditions across the whole area of the Netherlands.

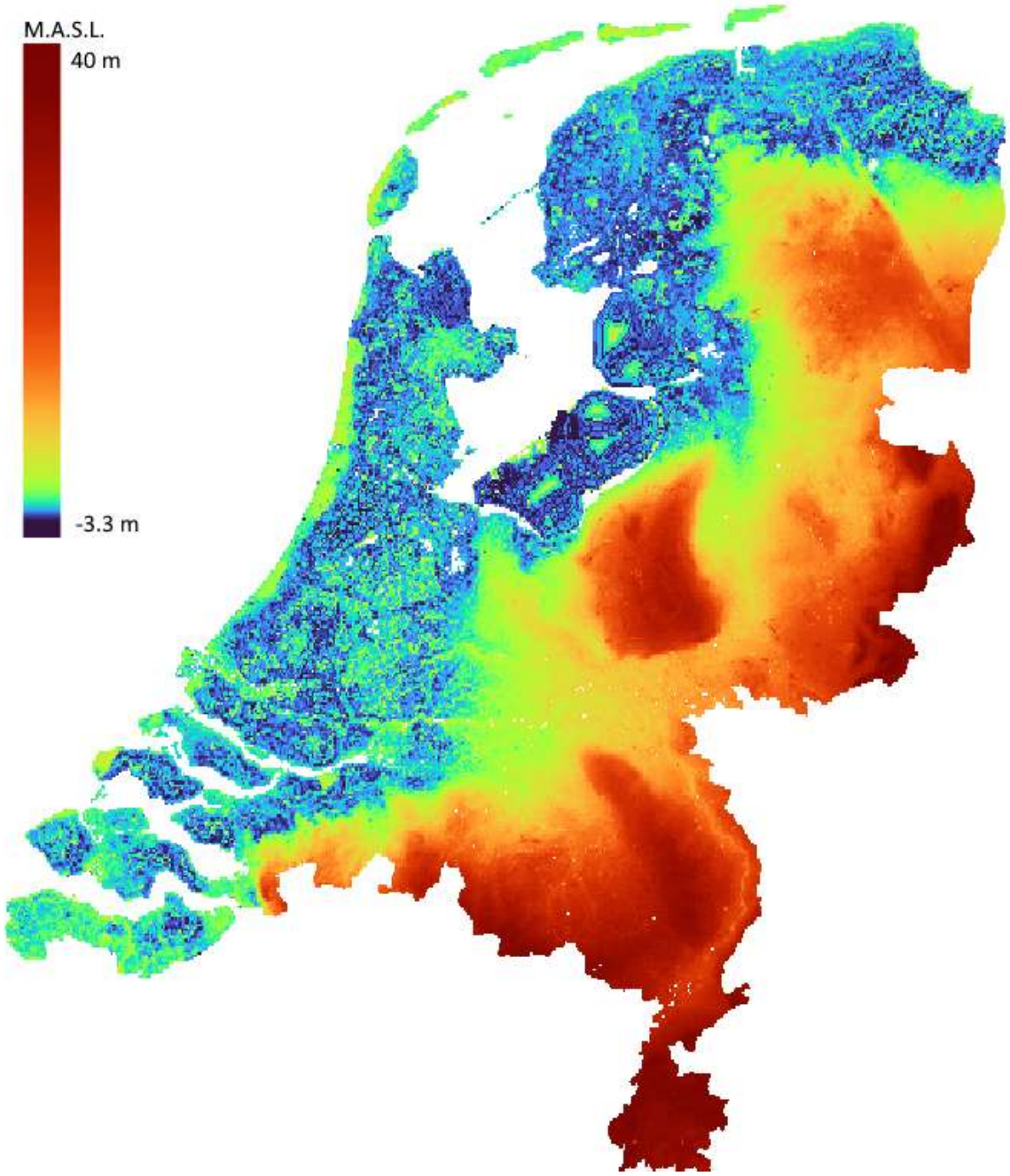



*Figure 29. Groundwater levels predicted by CB model for the whole area of the Netherlands. This quantile map represents the smaller scale variations of grounwater level depth in the country, especially in the northern parts with less deep groundwater levels*

# 4. Discussion

In this study the influence and significance of ML model selection and input datasets have been investigated to improve the performance and prediction skill of already existing groundwater level models. For this purpose, two different ML algorithms have been utilized with a considerable amount of input variables. The model incorporated the results of a nationwide numerical model. The input variables were chosen based on their potential relationship with groundwater level dynamics.

The selected variables reflect various meteorological, hydrological, and environmental processes, as well as general physical conditions. Prediction skills were assessed and compared by using different evaluation metrics and the importance of distinct features were calculated in different scenarios to investigate the possibilities of spatial dependency. Additionally, a groundwater level map was created for the whole Netherlands to represent the enhancement of groundwater level prediction quality. Finally, an additional model was developed to check whether there is a possibility to predict the error between groundwater level simulation results and actual groundwater level observations. Developing such models and understanding the potential relationship between variables could be an aid for future groundwater management and can be applied for further research to enhance groundwater level predictions in order to minimize the negative effects of possible future groundwater level extremes.

## 4.1 Potential of ML models for groundwater level predictions

One of the main objectives of this research was to show the potential of ML methods in predicting groundwater levels. A generalized model was constructed, tested, and validated for this purpose, to estimate groundwater levels not only at well sites where the groundwater level is known, but also at unknown locations. Two distinct ML algorithms were used during the whole modelling process. The results of a previously developed numerical groundwater level model were also employed to ensure physical consistency and improve model performance. In order to show the potential of different ML algorithms the results of the process-based generalized ML model and the original numerical model were compared. The largest and smallest 5% of groundwater observations were deleted to better represent typical groundwater level conditions and to minimize erroneous and incorrect data rows. The results show that both methods (RF and CB) are capable of predicting groundwater level dynamics. The algorithms were able to increase the numerical model's performance and correct large errors. The performance of the CB model is somewhat better than that of the RF model, however the CB model's computing time is significantly longer. Hyperparameter tuning was carried out for both methods, yet a complete grid search was only utilized for the CB model because of computational issues. In most circumstances, a complete grid search produces better results and model performance than a randomized search, hence it can be the main reason for the underperformance of the RF model. Given this information, it is reasonable to conclude that both implemented ML algorithms are adequate for such modelling applications.

The drawbacks of a generalized (groundwater level) model, as well as ML models in general, cannot be ignored. Although all ML models are dependent on large amounts of data, they are unable to detect physically incorrect data combinations, making data quality crucial for such algorithms. In addition, ML models may develop misleading associations during training if the input variables are invalid at some points or the numerical model results have a high inaccuracy at certain locations. In this case, these characteristics can be explained by the effect of two distinct attributes: 1) The results of the numerical model are already shifted in some degree due to modelling errors, and since this is the most important feature for both ML models the results will be close to these values. 2) During the modelling several features were used which were either categorical (i.e., soil type or land use type) or fixed values (i.e., elevation or distance from closest water body). Since these variables are not changing with time the weight of these values is increased. As a result, if the model learns during the training process that one or more of the constrained variables belong to a specific groundwater level value, predictions will be erroneous in cases where the groundwater value is significantly different, but the categorical or fixed variables are similar to the previous case. Data availability and quality are outstanding in the Netherlands however the aforementioned drawbacks can not be completely avoided, especially when employing the results of an existing model. Implementing the elimination of the highest and lowest 5% of the used groundwater level observation dataset significantly helped to attenuate these inaccuracies. As an example, *section 3.2.2* discussed that the majority of the large errors in the developed ML models (as well as in the employed numerical model) are spatially

dependent, and mostly occur in the southern regions of the country, where the groundwater levels are generally deeper. This spatial trend and dependency of error is further discussed in *section 4.3.*

## 4.2 Importance of input feature selection

The research also investigated the feature importance and relevance of the implemented input features to provide a broader picture of the modelling processes and present the relationships and possible connections between hydrological, meteorological, environmental variables and groundwater level fluctuations. Using this approach and examining how different variables influence model behaviour might help minimize the quantity of data required for calculation, reducing computational time and demand. The importance of various features was studied in different settings. First, the outputs of the two implemented ML models were compared to see whether there were any significant differences. Since the correlation between the used numerical model and actual groundwater level observations is significant, the importance of the model's results as features was substantially higher in every case, so it was removed from the comparison to better represent the importance of the remaining variables. According to the results there were no significant differences between the importance of the two ML models. In light of the findings, the categorical (land use type and soil type) and fixed (water distance and elevation) features are the most relevant. However, because these factors do not change over time, their weight in the comparison might be much larger than in reality. This is because a weekly time series of groundwater level observations (as well as meteorological and hydrological input data) were used with constant values, causing the algorithm to link these constraints to specific groundwater levels, lowering the valid feature importance of the implemented variables. To overcome this, a simpler model was created that does not account for temporal variability and thus only operates with constant values, therefore removing this difference. For the studied well locations, all of the data was averaged over the inspected time period to provide a single overall value for each input and target variable. The result of this simplified model shows that the most important variables are precipitation and evapotranspiration, next to the sea-level and water distance. This order is more in line with the anticipated outcomes of the importance calculations. Most groundwater level models (including numerical and ML methods as well) account for precipitation, evapotranspiration, discharge measurements of major river bodies and sea-level observations to represent tidal fluctuations. However, the majority of these models ignore the distance from the nearest water body (typically canals, rivers, and smaller lakes in the Netherlands), despite the fact that the findings of this model demonstrate that including it might improve prediction abilities and model performance in general.

Additionally, the dataset was separated into low and high lying areas considering the elevation. With this approach it might be possible to examine the distribution of feature relevance at different elevations. Firstly, the original model has been evaluated to see how the importance differs compared to the general model including all the data. Secondly, similarly to the previous approach, the feature importance was also calculated with the simplified model, to see how they behave when the temporal variability is excluded. According to the calculations, the importance of the generalized model has not been significantly changed. As it was discussed previously, this might be the result of the existence and overcalculated importance of the categorical and fixed variables. The simplified model might give a better and more realistic representation of the true feature importance across the country, including the low and high lying areas as well. For the simplified model's feature importance only the results of the RF model are presented in the paper, since the results of the CB model are considerably worse (for this scenario), thereby potentially not representing realistic values. These results are showing, that for low lying areas the four most important feature is elevation, evapotranspiration, sea-level and precipitation. Evapotranspiration and precipitation can be

considered as critical features, as their changes have a significant impact on groundwater recharge and hence groundwater dynamics in general. The tidal fluctuations can have a considerable impact on the northern water bodies (and groundwater) of the Netherlands, hence mostly important for the low-lying areas. These variations (especially during times of dry periods and droughts) have to be included in such models, since many of the northern regions water bodies are playing crucial roles in shipping, transit, and agricultural purposes as well. According to the results of the simplified model the most important feature is elevation. Clearly, modelling on a larger scale requires considering altitude differences. Since the groundwater level readings in these locations are often lower than in high-lying places, modest elevation variances can lead to significant differences in groundwater levels. For the southern part of the country, consisting of mainly high lying areas the most important features are precipitation, distance from the closest water body, the discharge of the Rhine River and evapotranspiration, respectively. In general, high parts are mostly dependent on precipitation regarding the groundwater recharge and less on recharge from rivers. However, since these areas are mainly located in the southern and eastern parts of the country the discharge of the Rhine River could also have significant effects on groundwater level fluctuations. In this example, and in the previously described scenarios of feature significance estimates in general, determining the distance from the nearest water body might be a critical component in estimating groundwater levels. These values were estimated with a 250m resolution in this model and had a considerable influence on groundwater levels in the majority of situations. Improving the precision of such models, and hence the accurate position of such water bodies, might make this feature even more valuable.

## 4.3 Investigating the location dependency of prediction errors

A comprehensive spatial analysis is necessary to create a realistic generalized groundwater level model and assess potential limits. The spatial variability of the error (in this example, the MSE values were used) and other assessment metrics may be analysed regionally using such an approach, hence different areas may be distinguished and the model's performance capabilities for distinct geographical locations can be assessed.

In order to adequately represent the spatial variability and potential trend in prediction error, the calculated MSE values for all the investigated well locations were visualized. The created point and heatmap showed that the average error in the model is significantly higher in the southern regions of the country, where the groundwater levels are potentially larger. The reduced amount of data for larger groundwater levels might be one of the reasons for this regional tendency. Because the majority of well observations are near to the median groundwater level, which in the Netherlands is often low, the (extremely) high water levels are not well represented, implying a higher risk of modelling error. Additionally, the 5% of these extremely high values were eliminated to improve the model performance, which can also be a potential cause of higher error values. Furthermore, the inaccuracies between the developed numerical model and real groundwater level measurements were generally higher for bigger groundwater levels on several instances. Both ML models were capable of lowering these errors, but they were unable to entirely eliminate them, hence, the inaccuracies of the utilized numerical model could be another major reason for the larger errors in southern locations with deeper groundwater levels. For large groundwater levels, the produced timeseries support the ML models' error-reducing characteristic. The ML models are more sensitive to weekly, smaller-scale fluctuations and the utilized variables are capable of reducing the magnitude of error in several cases (i.e.: Eindhoven or Maastricht case visualized in *Fig. 19*.).

Considering the given results, it can be concluded that the developed ML models are capable of lowering the errors of the numerical model, however still producing significant errors when the used groundwater level simulations values are highly inaccurate. This is mostly typical in the southern

and eastern parts of the Netherlands, where the groundwater levels are deeper, thereby the possibility for more significant errors is greater. The imbalance in the dataset is a potential problem and explanation for the substantial inaccuracies for deep groundwater levels. This indicates that the majority of the data belongs to one class (lower groundwater levels), while the minority of the data belongs to another (higher groundwater levels), resulting in a data distribution imbalance. When working with unbalanced datasets, the difficulty is that most machine learning approaches will overlook the minority class, resulting in poor performance. A possible way for future improvement is to oversample the minority class. This technique is not adding any new information; however, it might enhance the performance and predictive skills of the model.

## 4.4 Potential of ML models in developing error predictions

In addition to the initial model (and the many scenarios) outlined in section 4.1, an error model was created. The primary purpose of this case was to develop a model that could estimate differences (errors) between the existing groundwater level simulation model and actual groundwater level data. After a successful validation, such a model may be used to predict errors and, as a result, compute real groundwater levels without the need for any observations. However, the findings reveal that given the existing data and parameters, such a model is unable to find any connections between the input features and the target variable (i.e.: error). The model usually predicts a value around zero, which is mainly correct (as the bulk of actual error values are close to zero), but it also predicts a small error when the error value is significant in reality. Possible reasons for this can be that the model may be unable to pick up any relevant information and relationships, so providing no value and creating no differences in the outcomes, or that the majority of the errors are tightly distributed near zero, preventing the model from obtaining any helpful information for the minority class (i.e.: large errors corresponding to deeper groundwater levels). Two different techniques were used to overcome this problem. In both situations, the simplified, averaged model was applied, in which all the data was averaged into a single variable over the whole investigated time period, primarily to minimize computing time and demand. Firstly, the extreme data was removed: before training the model, 15% of the extremely low and high groundwater level observations, as well as the relevant input variables, were excluded. With this approach the results showed a significant improvement. Secondly, the greater error values were oversampled using a package called SMOGN to boost the model's prediction skills even further. This method allows the creation of a dataset that is closer to the normal distribution, making it easier to generate accurate predictions for the minority class as well. However, as compared to the prior method, this method produced higher evaluation errors. For larger errors and groundwater levels, it was able to provide more accurate estimations, yet the prediction skill decreased for values near zero.

In summary, it can be concluded that picking up correlations and connections between variables for error prediction is more difficult for such ML techniques than it is for groundwater level prediction. The significant correlation of the simulations with the groundwater measurements is not necessarily relevant for the errors, because the error computation requires subtracting the simulated and real groundwater levels. However, with different simplifications and additional techniques it is possible to develop a model which can be used as a starting point for the development of a similar model which might be validated for the initial model including the timeseries. Such a validated model would be a possible way to overcome the problems of unavailable and erroneous observational data and therefore simplifies the estimation of groundwater levels.

## 4.5 Limitations and possibility for further improvement

Aside from the uncertainties discussed and mentioned throughout the paper, there are numerous other uncertainties and limitations in this model. Two different ML approaches were

implemented and proved to be able of enhancing the quality of an existing numerical model, and therefore improve groundwater level estimations in general. All the calculations were carried out in a 250m resolution, which is too coarse for the accurate modelling of such hydrological processes. This resolution may be accurate for hydrological and meteorological measurements, although for the rest of the data (soil type, land use type, elevation, and distance from the closest water body), higher resolution would be required to provide more exact estimates (for example Koch et al., 2021). Additionally, it has been assumed that the discharge of the two rivers (Rhine and Meuse), as well as the sea-level measurements (Haringsvliet) have the same values for every location across the Netherlands. The model was capable to calculate the feature importance of these input variables and thereby give weight to their relevance, although the implementation of regional observations from different measurement points at more, spatially scattered locations or additional water bodies (e.g.: IJssel or De Lek rivers) could enhance the performance of the model and could give more accurate results for the actual feature importance as well. Furthermore, based on the availability and quality of observations, 13 separate meteorological regions were established. However, because of the irregular weather patterns in the Netherlands, additional meteorological sites, and therefore more unique meteorological regions might be employed, potentially improving the model's capabilities.

Furthermore, the utilization of timeseries and constant values together can create misleading, shifted results and errors in the feature importance and relevance of the input variables. Some data pairings may be contradictory because the utilized data may be incorrect in certain cases or because the applied resolution is too coarse. As an example, the distance from the closest water body was calculated for every pixel. For the whole timeseries, these computations were used for every cell with the same value. On some occasions it might happen that for very similar distances completely different groundwater level measurements are associated (e.g.: same distance from water body in the southern and northern parts of the Netherlands does not necessarily mean similar water levels, mainly because of differences in the elevation). Thereby, the model might learn misleading relationships during training and could make the model shifted (*Fig. D/1.*).

## 4.6 General summary and relation to current research

In general, the constructed model(s) may provide accurate and reliable predictions. They showed that the implemented ML algorithms (RF and CB) are capable of predicting groundwater levels as it was demonstrated by Koch et al. (2021), Hauswirth et al. (2021) and Wang et al. (2018). A potential improvement might be the utilization of other ML methods or the usage of neural networks for groundwater level modelling or prediction, such as LSTM's (Wunsch et al., 2021) or NARX's (Di Nunno et al., 2020). The research also supports the findings of Koch et al. (2021) where the authors revealed that by utilizing knowledge (physics) driven ML techniques, it is possible to precisely estimate groundwater levels with exceptional spatial accuracy. The findings of this study are in line with the Sahu et al. (2020), who showed that precipitation and river flow are relevant characteristics in many, but not all locations, nonetheless, creating reliable forecasts using only temperature and historical groundwater level data is insufficient. Meteorological features are critical to properly model groundwater level dynamics, although additional datasets (i.e., distance from the closest water body) could improve the predictions skills and quality of such models. In the following years further research is required to enhance the performance of such models. However, there are numerous ways to further increase prediction skill and overall performance. This improvement could be the implementation of and testing of different ML and neural network algorithms, introducing new, previously unused datasets as input variables or enhancing the resolution as much as the computational power makes it possible. In summary, the developed generic model may be used as a foundation for more refined and coarse groundwater level simulations, and hence could be a useful tool for future models.

# 5. Conclusion

In this research the potential of ML model and input feature selection was explored in order to develop an ML model what also incorporates physical consistency, and capable of improving already existing groundwater level simulations and helps to better understand relationships and connections between hydrological, meteorological, and environmental variables. Two distinct ML methodologies were used (namely RF and CB) to build, validate and test a generalized model what can reliably create groundwater level estimates. To evaluate the performance of the model, different metrics were used such as the MSE and RMSE values as well as the Pearson's correlation. In addition, the feature importance was computed for many cases to see how significant the implemented input variables are for groundwater level prediction and to see if there is any probable geographical pattern. Furthermore, an ML model was constructed to investigate if the difference between groundwater level simulations (results of the numerical model) and actual groundwater level measurements could be estimated. To better explain and visualize the findings a nationwide groundwater level map was created to show the average groundwater level conditions in the Netherlands.

Based on the findings, it can be determined that the developed ML algorithms are capable of improving the outcomes of the existing numerical model, and thus for groundwater level prediction in general. Both approaches performed similarly, although RF models are suggested for this purpose due to the necessary computing time. The generated model operates with much bigger errors for these locations as well, owing to the numerical model's larger inaccuracies for places with deeper groundwater levels (primarily the southern part of the Netherlands), although the model still improves the performance of the implemented numerical model. To further understand how feature relevance works in general, several scenarios were examined. It can be safely concluded that, despite some of the drawbacks, the incorporation of meteorological variables (precipitation and evapotranspiration) and physical variables (distance from the closest water body and elevation) are crucial for modelling such hydrological processes. The importance of hydrological factors must also be acknowledged; however, because these numbers were averaged over the whole country, their accuracy cannot be guaranteed. Finally, the error modelling findings reveal that estimating such differences where there is no actual association between the input and target variables is difficult. Due to simplification, averaging and reducing the dataset to exclude groundwater level extremes an error model was developed with promising results. With additional refinement, such as the addition of new variables, correction of data sample disparities, or transformation of some variables, this model might be a useful starting point for future study.

In conclusion, such ML algorithms may be employed not only for simulating current or historical observations, but also for forecasting. Despite the shortcomings and limitations of the created models, these approaches have a lot of potential, but also a lot of room for improvement in terms of forecasting groundwater levels and other hydrological processes in general. The findings can be used as a base and starting point in future research to improve groundwater level predictions and, as a result, water management strategies in order to reduce the damaging effects of future groundwater level extremes that could result in severe droughts or floods.

# Appendix A.

*Land use map and classes of the Netherlands*

The employed land use dataset contains 39 instinct classes (*Tab. A/1*). *Fig. A/1.* shows the distinct classes which represents the different land usage categories all across the Netherlands.

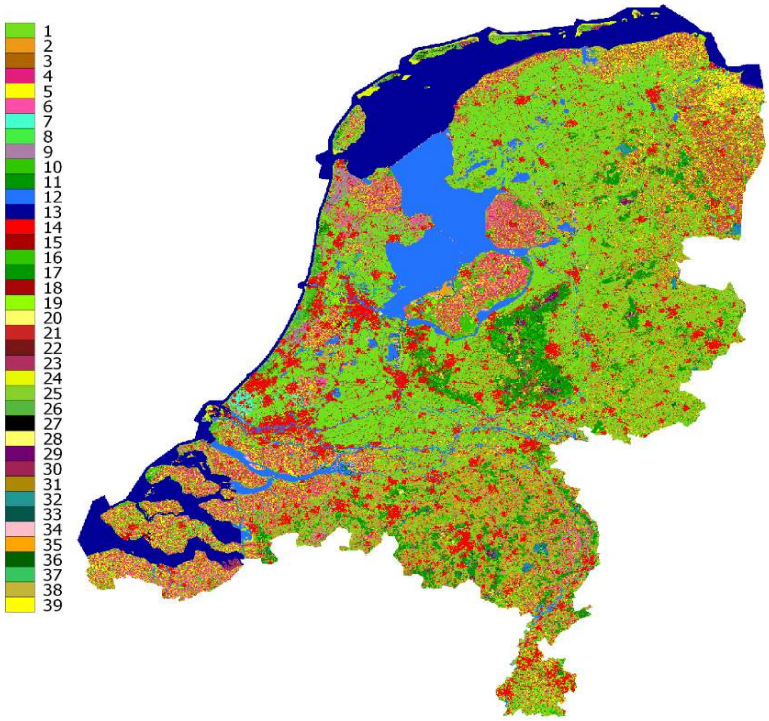



*Figure A/1. Land use map and classes of the Netherlands (LGN4 model). The explanation of the classes can be found in Table A/1.*

To keep the model simple and robust, the different land use categories were converted into 16 different categories by considering the type of land usage (in many cases similar classes were separated) and their frequency. *Tab. A/2.* Includes the narrowed categorical classes of land use type.

| Class | Land use | Class | Land use | Class | Land use | Class | Land use |
|---|---|---|---|---|---|---|---|
| 1 | Grassland | 11 | Coniferous forest | 21 | Main roads and rail | 31 | Heavily grassed heather |
| 2 | Corn | 12 | Saltwater | 22 | Buildings in agricultural areas | 32 | Raised bogs |
| 3 | Potato | 13 | Freshwater | 23 | Salt marshes | 33 | Forest in moorlands |
| 4 | Beets | 14 | Urban built-up area | 24 | Open sand in coastal area | 34 | Other swamp vegetation |
| 5 | Cereal | 15 | Buildings in rural areas | 25 | Open dune vegetation | 35 | Reed vegetation |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 6 | Other agricultural crops | 16 | Deciduous forest in built-up areas | 26 | Closed dune vegetation | 36 | Forest in swamp vegetation |
| 7 | Greenhouse horticulture | 17 | Coniferous forest in built-up areas | 27 | Dune | 37 | Peat meadow area |
| 8 | Orchard | 18 | Densely built-up forest | 28 | Open drifting sand | 38 | Other |
| 9 | Flower fields | 19 | Grass in built-up areas | 29 | Heather | 39 | Kale ground |
| 10 | Deciduous forest | 20 | Bare ground in built-up areas | 30 | Moderately grassed heather | | |

*Table A/1. Original land use classes of LGN4*

| Class | Land use | Gathered classes | Class | Land use | Gathered classes |
|---|---|---|---|---|---|
| **1** | Grassland | 1, 19 | 9 | Salt marshes | 23 |
| **2** | Agricultural crops | 2, 3, 4, 5, 6, 7, 8 | 10 | Sand | 24, 28 |
| **3** | Flower fields | 9 | 11 | Dune vegetation | 25, 26, 27 |
| **4** | Deciduous forest | 10, 16, 33, 36 | 12 | Heather | 29, 30, 31 |
| **5** | Coniferous forest | 11, 17 | 13 | Raised bogs and swamp vegetation | 32, 34 |
| **6** | Water | 12, 13 | 14 | Reed | 35 |
| **7** | Buildings and roads | 14, 15, 21, 22 | 15 | Peat | 37 |
| **8** | Bare ground | 20, 39 | 16 | Other | 38 |

*Table A/2. Compressed land use classes. These categorical variables were used during modelling*

# Appendix B.

*General overview and visualization of the used datasets - Distribution of wells in different meteorological regions and some of the employed data (precipitation, evapotranspiration, sea-level, discharge of Rhine and Meuse, soil type, land use type and water distance). The goal of this appendix is to provide insight and thereby a better understanding of the used data.*

*Fig. B/1.* shows the distribution of wells in different investigated meteorological regions. The regions were chosen based on data availability (in the investigated time period between 1980 and 2019) and dispersity to cover different areas. However, as seen in *Fig. B/1.*, wells are not evenly distributed among the various areas. This is due to the different areas of the regions, as well as the spread of wells (the well density close to the coast and in the northern areas is considerably lower compared to the middle and southern parts of the country).

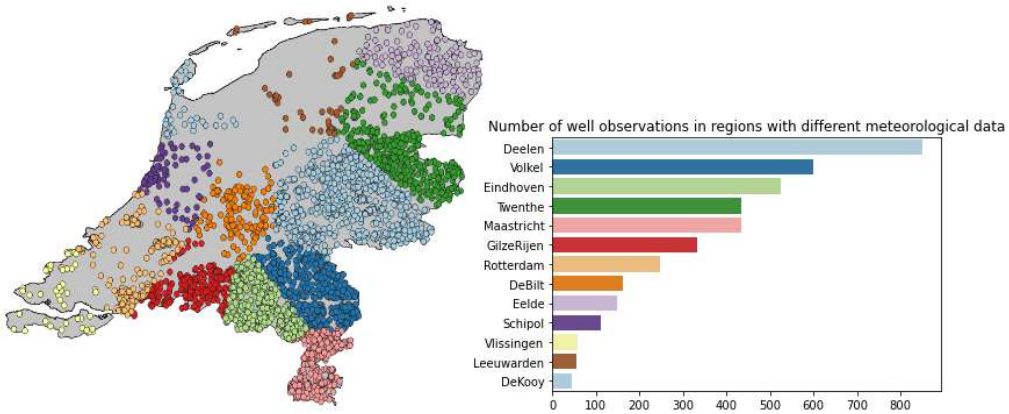



*Figure B/1. Locations of the investigated well and their frequency by regions with different meteorological data*

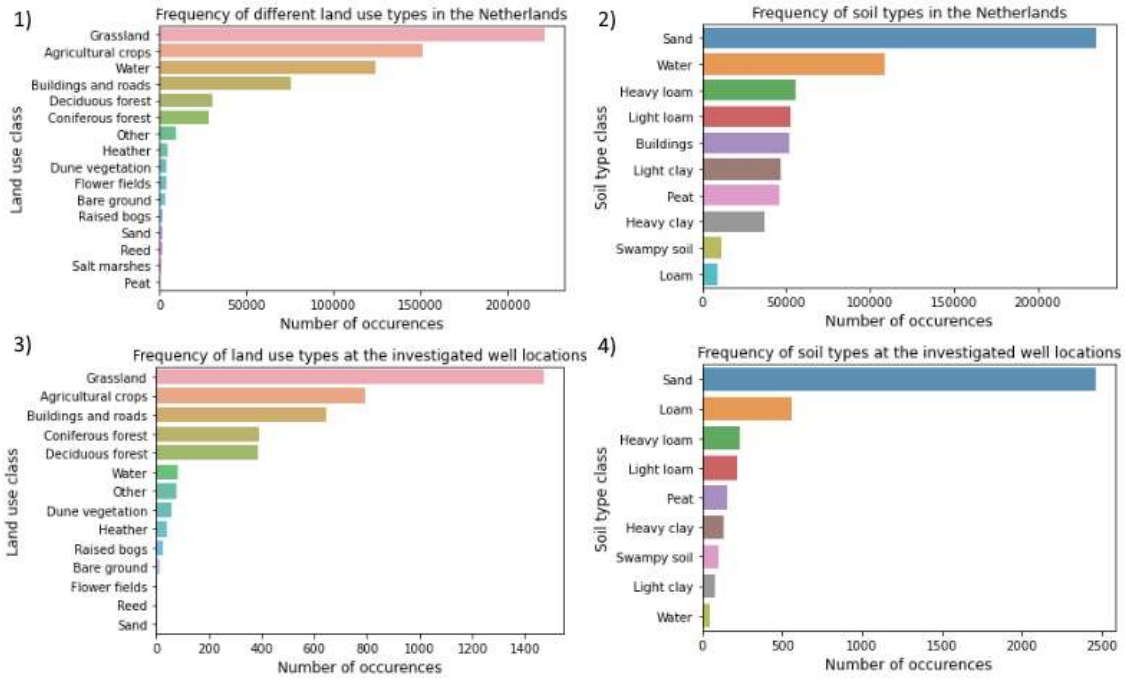



*Figure B/2. Frequency of different land use and soil types across the country and at the investigated well locations*

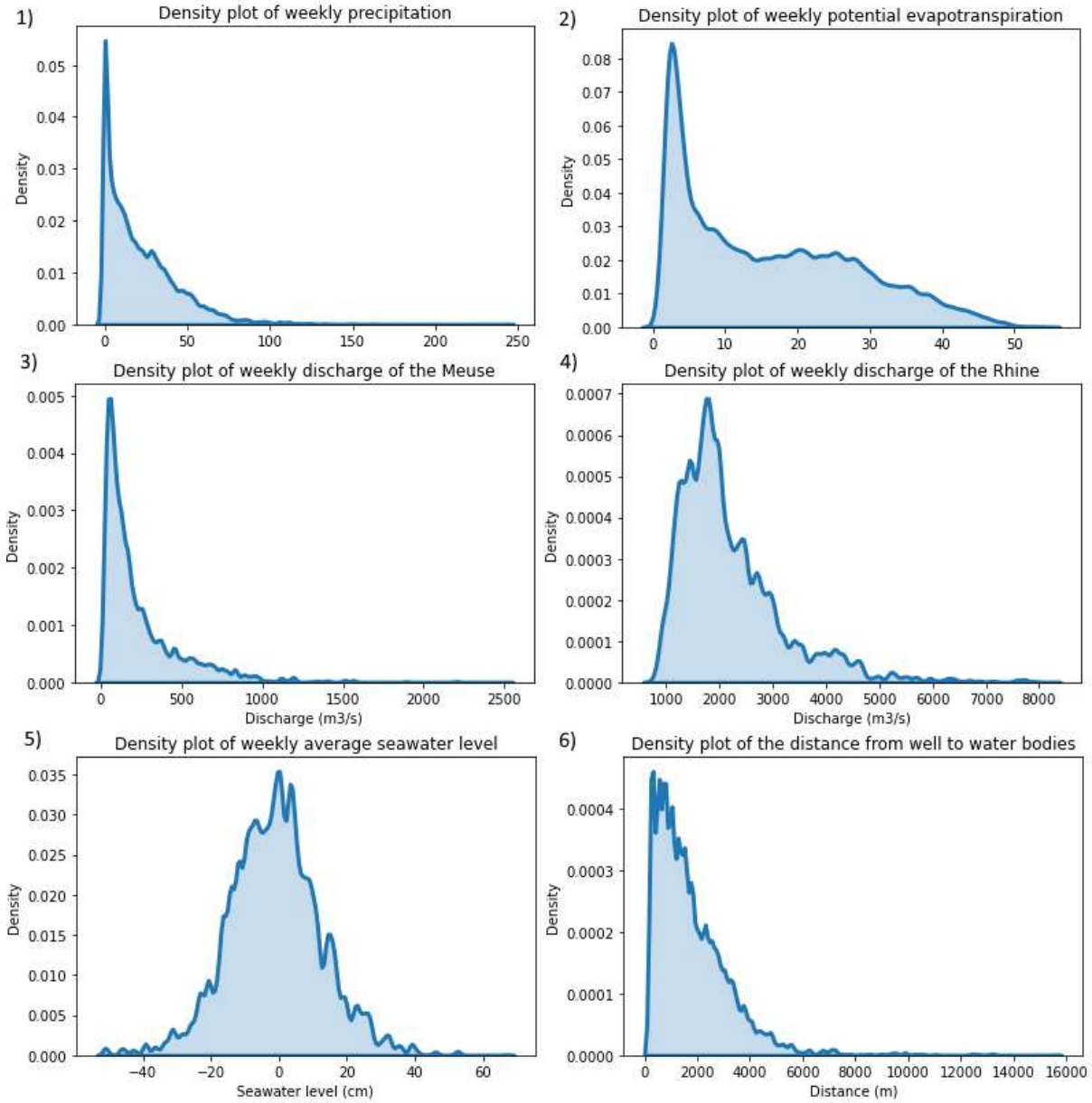



*Figure B/3. Density plots of different input variables. 1) Precipitation, 2) Potential evapotranspiration, 3) Discharge of the river Meuse, 4) Discharge of the river Rhine, 5) Seawater level fluctuations due to tidal effects at Haringsvliet, 6) Closest distance to any type of waterbody in a 250-meter resolution*

*Fig. B/2* and *B/3.* depicts a general overview of the implemented input variables. *Fig. B/2.* gives a visualization about the used soil and land use type data and the frequency of different values contained in these datasets. *Fig. B/3.* shows the distribution of hydrological and meteorological observations.

# Appendix C.

*Exact feature importance values for both RF and CB models (including the original case, considering groundwater simulations as well and the reduced case, excluding the simulations).*

| Features | Original RF (%) | Original CB (%) | Reduced RF (%) | Reduced CB (%) |
|---|---|---|---|---|
| Simulations | 97.542874 | 76.500491 | - | - |
| Elevation | 1.082923 | 9.289539 | 44.072728 | 39.530779 |
| Water distance | 0.944868 | 7.624736 | 38.454204 | 32.446365 |
| Land use type | 0.156971 | 3.265740 | 6.388402 | 13.897058 |
| Soil type | 0.126983 | 2.110490 | 5.167940 | 8.980997 |
| Discharge of Meuse | 0.044027 | 0.313500 | 1.791803 | 1.334069 |
| Evapotranspiration | 0.034790 | 0.263714 | 1.415902 | 1.122211 |
| Discharge of Rhine | 0.025750 | 0.240123 | 1.047992 | 1.021821 |
| Seawater-level | 0.022933 | 0.241163 | 0.933328 | 1.026246 |
| Precipitation | 0.017881 | 0.150503 | 0.727701 | 0.640453 |

*Figure C/1. Feature importance of both RF and CB models, including and excluding the groundwater level simulations (original case).*

| Case | Importance (%) |
|---|---|
| Low RF | 88.137 |
| Low CB | 91.044 |
| High RF | 92.178 |
| High CB | 85.882 |

*Figure C/2. Feature importance of groundwater level simulations both RF and CB models, for low and high lying areas*

| Features | RF case low | CB case low | RF case high | CB case high |
|---|---|---|---|---|
| Precipitation | 0.126 | 0.069 | 0.268 | 0.118 |
| Evapotranspiration | 0.184 | 0.259 | 0.457 | 0.272 |
| Seawater-level | 0.164 | 0.156 | 0.402 | 0.198 |
| Discharge of Meuse | 0.283 | 0.459 | 0.511 | 0.417 |
| Discarge of Rhine | 0.191 | 0.199 | 0.312 | 0.316 |
| Land use type | 12.572 | 12.059 | 14.057 | 10.986 |
| Elevation | 40.643 | 33.218 | 37.475 | 37.121 |
| Water distance | 34.323 | 40.988 | 36.601 | 40.658 |
| Soil type | 11.515 | 12.592 | 9.916 | 9.912 |

*Figure C/3. Feature importance of both RF and CB models, for low and high lying areas, excluding the groundwater level simulations (original case).*

# Appendix D.

*Timeseries of the remaining randomly selected well locations*

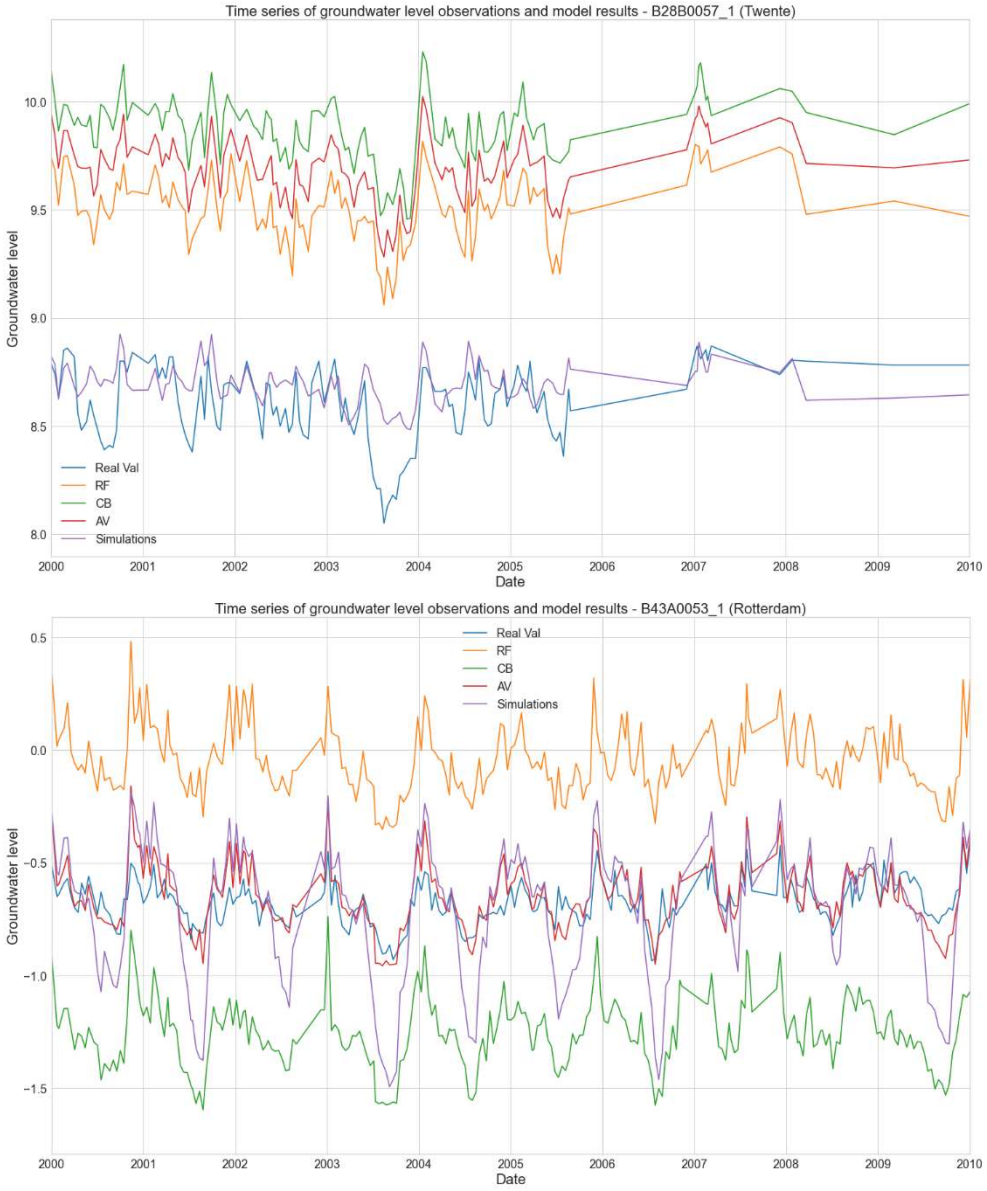



*Figure D/1. Time series of 2 different well locations. Up: Twente region (well ID: B28B0057_1); Down: Rotterdam region (well ID: B43A0053_1).*

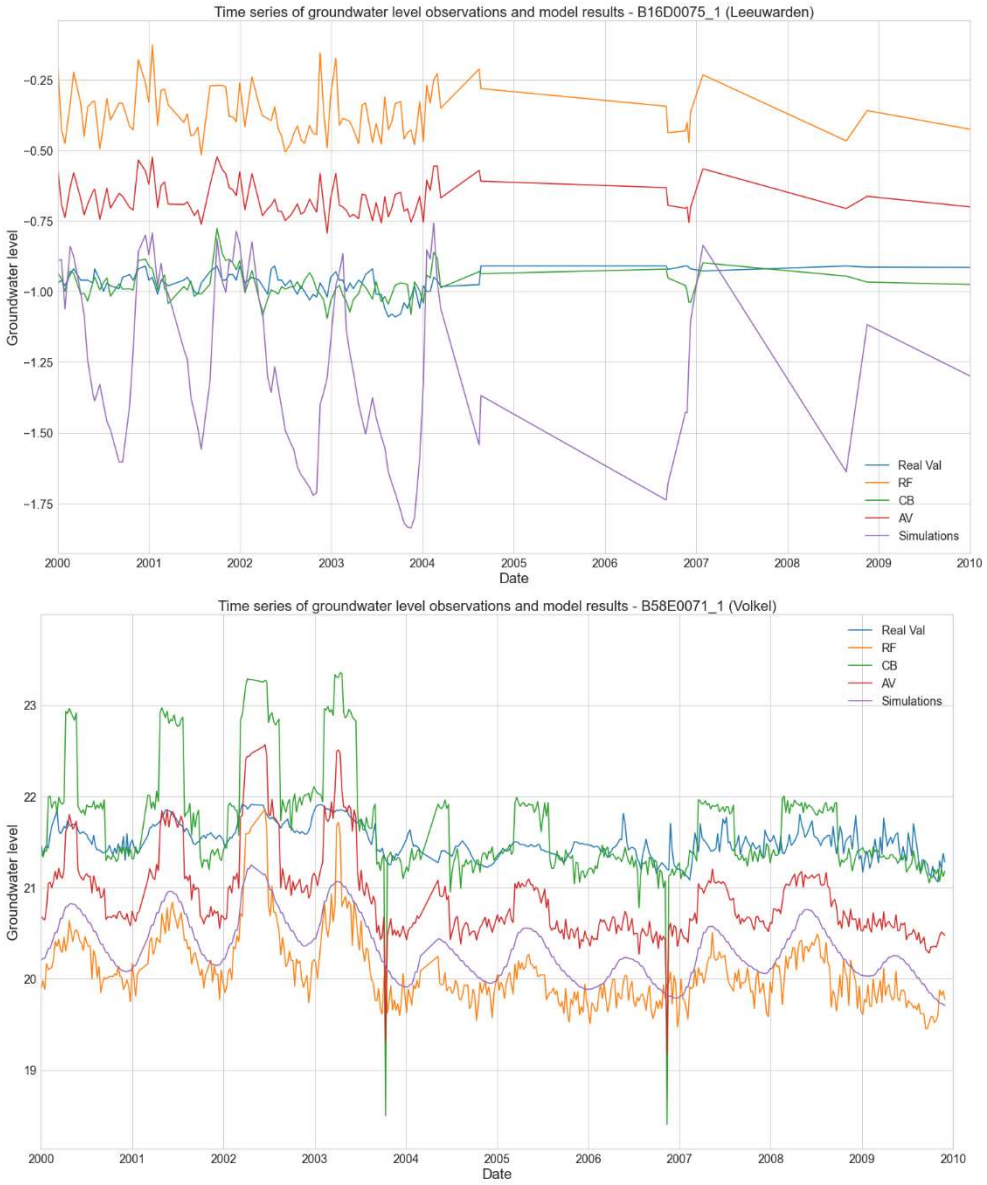



*Figure 18. Time series of 4 different well locations. Up: Leeuwarden region (well ID: B16D0075_1); Down: Volkel region (well ID: B58E0071_1)*

# Appendix E.

*Evaluation metrics of the investigated wells.*

| Well | MSE Sim | Pearsons Sim | MSE RF | Pearsons RF | MSE CB | Pearsons CB | MSE AV | Pearsons AV |
|---|---|---|---|---|---|---|---|---|
| Maastricht | 5.129 | 0.922 | 0.234 | 0.584 | 2.281 | 0.81 | 0.814 | 0.827 |
| Eindhoven | 0.172 | 0.668 | 0.068 | 0.658 | 0.722 | 0.701 | 0.178 | 0.727 |
| De Kooy | 0.0753 | 0.745 | 0.203 | 0.678 | 0.285 | 0.732 | 0.031 | 0.725 |
| De Bilt | 0.101 | 0.789 | 0.032 | 0.558 | 0.272 | 0.665 | 0.109 | 0.676 |
| Rotterdam | 0.061 | 0.763 | 0.405 | 0.742 | 0.342 | 0.733 | 0.009 | 0.784 |
| Twente | 0.0293 | 0.547 | 0.842 | 0.737 | 1.607 | 0.778 | 1.191 | 0.797 |
| Leeuwarden | 0.231 | 0.436 | 0.349 | 0.175 | 0.003 | 0.259 | 0.089 | 0.242 |
| Volkel | 1.298 | 0.799 | 1.917 | 0.736 | 0.335 | 0.702 | 0.469 | 0.743 |

*Figure D/1. Evaluation metrics for the randomly selected well locations*

# Appendix F.

*Scatter and residual plots of different scenarios of the error model*

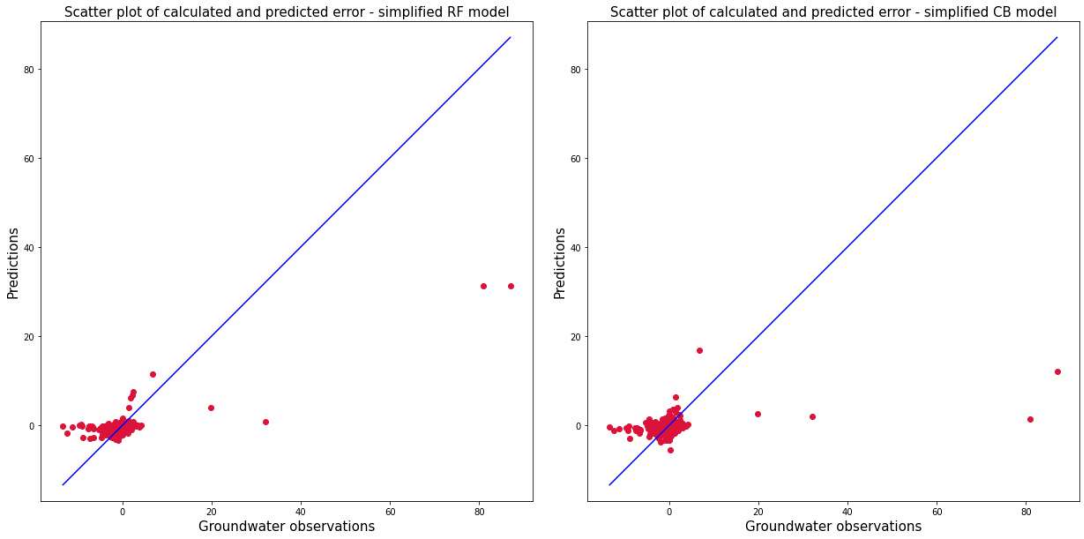



*Figure F/1. Scatter plot of the simplified error model in Section 3.3.2*

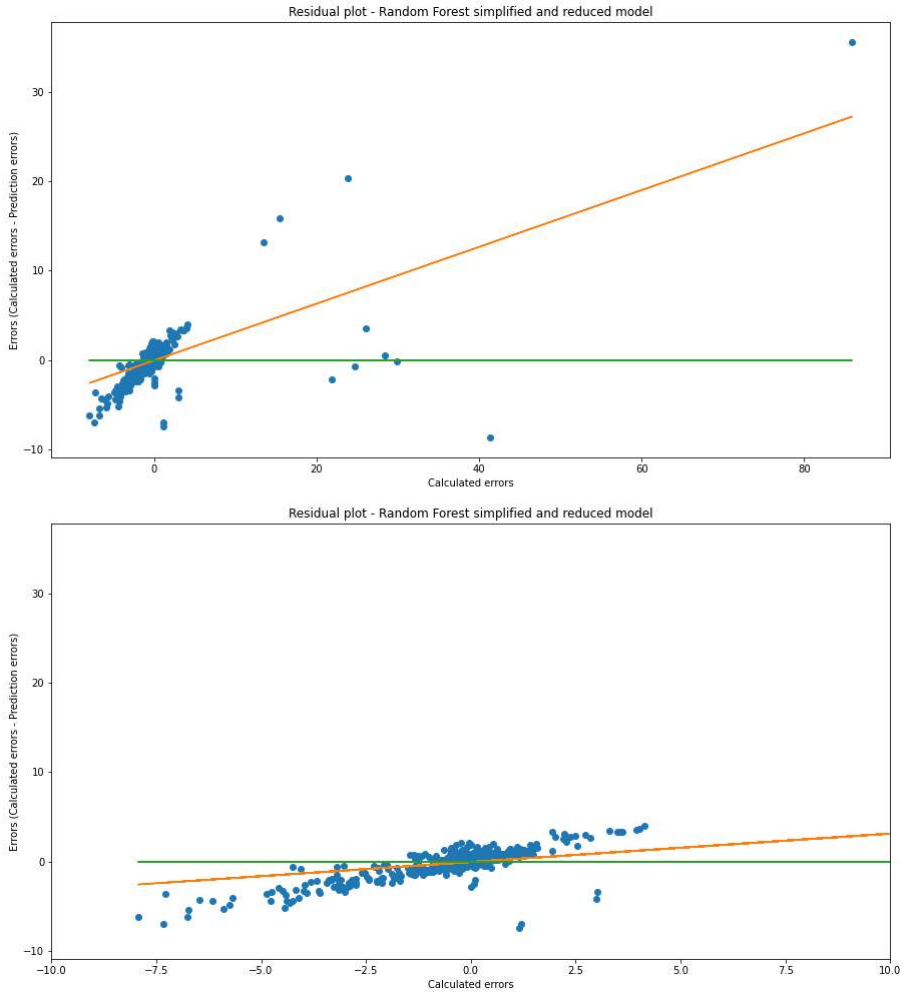



*Figure F/2. Residual plot of the simplified error model in Section 3.3.2*

# References


Acero Triana, J.S.; Chu, M.L.; Guzman, J.A.; Moriasi, D.N.; Steiner, J.L., (2020). Evaluating the Risks of Groundwater Extraction in an Agricultural Landscape under Different Climate Projections, *Water* **2020**, *12*, 400. https://doi.org/10.3390/w12020400

Allen-Dumas, M.R., Xu, H., Kurte, K.R. and Rastogi, D., (2021). Toward Urban Water Security: Broadening the Use of ML Methods for Mitigating Urban Water Hazards., *Front. Water* 2:562304. https://doi.org/10.3389/frwa.2020.562304

Alley, William & Reilly, T.E. & Franke, O.L., (1999). Sustainability of Ground-Water Resources. *U.S. Geological Survey Circular* 1186.

Asher, M. J., Croke, B. F. W., Jakeman, A. J., and Peeters, L. J. M., (2015) A review of surrogate models and their application to groundwater modelling, *Water Resour. Res.*, 51, 5957–5973. https://doi.org/10.1002/2015WR016967

Banerjee, P., Singh, V. S., Chatttopadhyay, K., Chandra, P. C., and Singh, B. (2011). Artificial neural network model as a potential alternative for groundwater salinity forecasting, *J. Hydrol*. 398, 212–220. https://doi.org/10.1016/j.jhydrol.2010.12.016

Bierkens, M.F.P., Wada, Y., (2019). Non-renewable groundwater use and groundwater depletion: a review, *Environ. Res. Lett.* **14** 063002. https://doi.org/10.1088/1748-9326/ab1a5f

Brakkee, E., van Huijgevoort, M., and Bartholomeus, R. P., (2021). Spatiotemporal development of the 2018–2019 groundwater drought in the Netherlands: a data-based approach, *Hydrol. Earth Syst. Sci. Discuss.* [preprint]. https://doi.org/10.5194/hess-2021-64

Branco, P., Torgo, L., Ribeiro, R. (2017). SMOGN: A Pre-Processing Approach for Imbalanced Regression. Proceedings of Machine Learning Research, 74:36-50.

Di Nunno, F., Granata, F., (2020). Groundwater level prediction in Apulia region (Southern Italy) using NARX neural network, *Environ Res*. 190:110062. https://doi.org/10.1016/j.envres.2020.110062

Gleeson, T., Befus, K. M., Jasechko, S., Luijendijk, E., and Cardenas, M. B., (2016). The global volume and distribution of modern groundwater, *Nat. Geosci*. 9, 161–167. https://doi.org/10.1038/ngeo2590

Goodfellow, I., Bengio, Y., and Courville, A., (2016). Deep Learning, *The MIT Press.* https://www.deeplearningbook.org

FAO (Food and Agriculture Orgianization of the United Nations), (2021). New FAO report on land and water resources paints an alarming picture.

Hauswirth, S.M., Bierkens, M.F.P., Beijk, V., Wanders, N., (2021). The potential of data driven approaches for quantifying hydrological extremes*, Advances in Water Resources*, Volume 155, 2021, 104017, ISSN 0309-1708. https://doi.org/10.1016/j.advwatres.2021.104017

Klein Tank, A., Beersma, J., Bessembinder, J., Van den Hurk, B., Lenderink, G., (2014). KNMI'14 Climate Scenarios for The Netherlands, *Technical Report*, KNMI, De Bilt, The Netherlands. http://projects.knmi.nl/publications/showAbstract.php?id=10756

Knoll, L., Breuer, L., Bach, M., (2019). Large scale prediction of groundwater nitrate concentrations from spatial data using ML, *Science of The Total Environment*, Volume 668, 2019, Pages 1317-1327, ISSN 0048-9697. https://doi.org/10.1016/j.scitotenv.2019.03.045

Koch, J., Gotfredsen, J., Schneider, R., Troldborg, L., Stisen, S, Henriksen, H. J., (2021). High Resolution Water Table Modelling of the Shallow Groundwater Using a Knowledge-Guided Gradient Boosting Decision Tree Model, *Frontiers in Water*, volume 3, 2021, p. 81. https://doi.org/10.3389/frwa.2021.701726

Koch, J., Berger, H., Henriksen, H. J., and Sonnenborg, T. O., (2019). Modelling of the shallow water table at high spatial resolution using random forests, *Hydrol. Earth Syst. Sci*., 23, 4603–4619. https://doi.org/10.5194/hess-23-4603-2019

Kraft, B., Jung, M., Körner, M., and Reichstein, M., (2020). Hybrid Modelling: Fusion of a deep learning approach and a physics-based model for global hydrological modelling, *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.,* XLIII-B2-2020, 1537–1544. https://doi.org/10.5194/isprs-archives-XLIII-B2-2020-1537-2020

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A.K., Hochreiter, S., Nearing, G.S., (2019). Towards improved predictions in ungauged basins: exploiting the power of ML, *Water Resour Res*. https://doi.org/10.1029/2019WR026065

Kuhn, M., Johnson, K., (2013). Applied Predictive Modelling, *Springer New York*, NY, https://doi.org/10.1007/978-1-4614-6849-3

Lijzen, J. P. A., Otte, P., & van Dreumel, M., (2014). Towards sustainable management of groundwater: Policy developments in The Netherlands. *Science of the Total Environment,* 485-486(1), 804-809. https://doi.org/10.1016/j.scitotenv.2014.02.081

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., et al. (2020). What role does hydrological science play in the age of ML? *Water Resour. Res.* 57: e2020WR028091. https://doi.org/10.1029/2020WR028091

Meyal, A.Y., Versteeg, R., Alper, E., Johnson, D., Rodzianko, A., Franklin, M., and Wainwright, H., (2020). Automated Cloud Based Long Short-Term Memory Neural Network Based SWE Prediction, *Front. Water* 2:574917. https://doi.org/10.3389/frwa.2020.574917

Mosavi, A., Sajedi-Hosseini, F., Choubin, B., Taromideh, F., Rahi, G., Dineva, A.A., (2020). Susceptibility Mapping of Soil Water Erosion Using ML Models, *Water*. 2020, 12(7):1995. https://doi.org/10.3390/w12071995

Philip, S.Y., Kew, S.F., van der Wiel, K., Wanders, N., van Oldenborgh G.J., (2020). Regional differentiation in climate change induced drought trends in the Netherlands, *Environ. Res. Lett.* 15 094081. https://doi.org/10.1088/1748-9326/ab97ca

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., et al., (2019). Deep learning and process understanding for data-driven Earth system science, *Nature* 566, 195–204. https://doi.org/10.1038/s41586-019-0912-1

Sahoo, S., Russo, T.A., Elliott, J., and Foster, I., (2017). ML algorithms for modelling groundwater level changes in agricultural regions of the U.S., *Water Resour. Res.,* 53, 3878–3895. https://doi.org/10.1002/2016WR019933.

Sahu, R.K., Müller, J., Park, J., Varadharajan, C., Arora, B., Faybishenko, B., and Agarwal, D., (2020). Impact of Input Feature Selection on Groundwater Level Prediction from a Multi-Layer Perceptron Neural Network, *Front. Water* 2:573034. https://doi.org/10.3389/frwa.2020.573034

Shen, C., Chen, X., and Laloy, E., (2021). Editorial: Broadening the Use of ML in Hydrology. *Front. Water* 3:681023. https://doi.org/10.3389/frwa.2021.681023

Siebert, S., Burke, J., Faures, J.M., Frenken, K., Hoogeveen, J., Döll, P., et al. (2010). Groundwater use for irrigation - A global inventory. *Hydrol. Earth Syst. Sci.* 14, 1863–1880. https://doi.org/10.5194/hess-14-1863-2010

Stisen, S., Sonnenborg, T. O., Refsgaard, J. C., Koch, J., Bircher, S., and Jensen, K. H., (2018). Moving beyond runoff calibration – Multi constraint optimization of a surface subsurface-atmosphere model, *Hydrol. Process*., 32, 2654–2668. https://doi.org/10.1002/hyp.13177

Verkaik, J., Hughes, J.D., van Walsum, P.E.V., Oude Essink, G.H.P., Lin, H.X., Bierkens, M.F.P., (2021). Distributed memory parallel groundwater modelling for the Netherlands Hydrological Instrument, *Environmental*

*Modelling & Software*, Volume 143, 2021, 105092, ISSN 1364-8152. https://doi.org/10.1016/j.envsoft.2021.105092

Wada, Y., Wisser, D., Bierkens, M.F.P., (2014). Global modelling of withdrawal, allocation and consumptive use of surface water and groundwater resources, *Earth Syst. Dyn*. 5, 15–40. https://doi.org/10.5194/esd-5-15-2014.

Wang, X., Liu, T., Zheng, X., (2018). Short-term prediction of groundwater level using improved random forest regression with a combination of random features, *Appl Water Sci* 8, 125. https://doi.org/10.1007/s13201-018-0742-6

Wunsch, A., Liesch, T., Broda, S., (2021). Groundwater level forecasting with artificial neural networks: a comparison of long short-term memory (LSTM), convolutional neural networks (CNNs), and non-linear autoregressive networks with exogenous input (NARX), *Hydrol. Earth Syst. Sci.,* 25, 1674-1687. https://doi.org/10.5194/hess-25-1671-2021

Zhao, G., Bates, P., Neal, J., and Pang, B., (2021). Design flood estimation for global river networks based on ML models, *Hydrol. Earth Syst. Sci.,* 25, 5981–5999. https://doi.org/10.5194/hess-25-5981-2021