Detecting unwanted consequences of a decreased maintenance condition in the main waterways of drainage areas using predictive machine learning techniques

Master thesis written by Thijs Roos BSc

Thijs Roos BSc	4288815
Water Science and Management	30 ETCS
Utrecht University	Dr. Ir. N. Wanders & Dr. M. T. H. Van Vliet
Hoogheemraadschap De Stichtse Rijnlanden	T. De Lange MSc
February 2020	July 2020

Utrecht University

HOOGHEEMRAADSCHAP DE STICHTSE RIJNLANDEN

Abstract

Maintaining the condition of the drainage areas is an important task of the regional water authority “Hoogheemraadschap De Stichtse Rijnlanden”, abbreviated by HDSR. HDSR prefers to monitor the maintenance condition of the drainage areas continuously in order to keep the condition at an acceptable level. However, observing this condition continuously for each drainage area is financially unfeasible. The aim of the study is to develop an approach to measure the maintenance condition of the main waterways of the drainage areas of HDSR in real time. This approach requires the water level difference in the waterways, Δh, as a parameter of the maintenance condition, since Δh is linked to flow resistance. In order to develop the approach, two case studies were performed: one in the Amerongerwetering drainage area and one in the Lange Weide drainage area. During the study, machine learning techniques, such as a linear regression model, a random forest model and a gradient boosting model, were applied. The models required a large input dataset to predict the Δh values. These value were compared to the observed values of Δh. When observing a significant difference between the predicted and the observed values, the date was classified as an anomaly. The data included in the study were provided by HDSR and KNMI. The linear regression model was not suitable for the study, because of insufficient prediction quality. Both the results of the random forest and gradient boosting model showed that most of the anomalies were detected. The anomalies were analysed and were linked to possible explanations. This analysis explained that excessive vegetation has large influence on Δh. The approach proved more promising for the (simpler) Amerongerwetering drainage area, compared to the Lange Weide drainage area. The random forest model proved to be a better performing model, both statistically and visually. The study concludes that machine learning provides opportunities for the water management in the drainage areas of HDSR. However, it is recommended that these opportunities are further examined in future studies.

Table of Contents

1 Introduction

1.1 Background

Over time, the maintenance condition of the main water ways of drainage areas tend to get worse. Most often, the conditions decrease is a result of excessive vegetation and dreck that decrease the flow velocity, broken pumps that do not deliver the required capacity and obstructions in the streams. The regional water authority “Hoogheemraadschap De Stichtse Rijnlanden”, abbreviated with HDSR, does not have a real time view on the full maintenance condition, since it is financially unfeasible to observe the condition of each drainage area by hand. However, HDSR is interested in knowing the maintenance condition since it is responsible for the surface water in the area. Drainage areas with a decreased condition could lead to several different problems as improper functioning of the system, floods and overall, an increase in costs. An approach which could help with assessing the maintenance condition automatically and real time could decrease the monitoring costs for HDSR intensively and problems could be detected earlier.

The research takes place at the regional water authority HDSR in the form of an internship. The focus of the research is on the development and testing of an approach to detect the maintenance condition of the main waterways of the drainage areas with the help of machine learning and anomaly detection techniques. After the development of the approach, it is tested in two different drainage areas. These tests will be described in the form of case studies.

By conducting two case studies, the approach will be tested for both a simple and a complex situation. The first case study is conducted in the Amerongerwetering drainage area, which has a relatively simple hydrology and land use. These factors make the drainage area perfect for a first test case for the methodology. The second case study is performed in the Lange Weide drainage area. Because of the complex hydrology and frequently changes over time, this drainage area is perfect for testing the capability of using the approach in such a situation. The relation between these two drainage areas could help with developing a more generic approach for more drainage systems. The locations of the two case studies and the total area that is managed by HDSR are highlighted in figure 1.1.

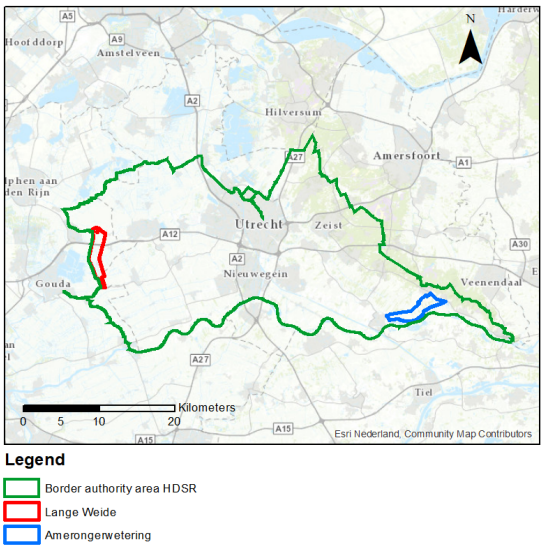

Figure 1.1: The area that is managed by HDSR and the location of the Amerongerwetering and Lange Weide drainage areas.

1.2 Problem description

The problem to be faced during this study is the issue that monitoring the maintenance condition of drainage areas real time is financially unfeasible, when done manually. This study focusses on the development of a new approach using anomaly detection and machine learning techniques on the stream gradient within the drainage areas. Resulting in a continuous view on the maintenance condition of the main waterways of the drainage areas of HDSR.

It is important to study whether there is a correlation between the gradient and irregularities in the drainage area, since this correlation could be used for various means, like finding irregularities in the, and assessing the condition of the drainage area remotely. Studying this area can also argued to be relevant because of the following reasons.

At first, the research topic is scientifically relevant since this is the first study on this specific topic. There is a knowledge gap present, as there is no literature which combines machine learning techniques and stream gradients in drainage areas. Most literature is focussed on the relationship between resistance and the stream gradient in high-gradient streams (Yochum & Bledsoe, 2010).

In addition, the societal relevance of the research is comprehensive, since the new approach could be used in order to increase the maintenance condition of the streams. Overall the drainage area could be working more efficient after the implementation of the method, so less problems are to be expected. Improving the efficiency of the regional water authority could save public capital, since regional water authorities are governmental organisations, financed by public funds. In this way, the approach could be advantageous for both HDSR as other regional water authorities in The Netherlands.

At last, the societal relevance has an overlap with the specific relevance for HDSR, since the regional water authority is interested in knowing the maintenance condition of the main waterways of the drainage areas. HDSR does not have a real time view on the maintenance condition of the drainage areas since it is financially unfeasible to observe the condition of each separate location continuously. However, an alternative approach of monitoring the maintenance condition could be interesting for HDSR. In particular if the approach is able to measure the condition automatically and real time. Potential changes in maintenance condition could be detected earlier by the new approach compared to the use of the current approach, due to the continuous measurement and a lower threshold. As a result, a more efficient maintenance schedule of the drainage areas could be designed. Furthermore, the approach has the potential to decrease the monitoring costs significantly, since the required data is already measured.

1.3 Literature review

Since flow resistance is proportionate with gradient and disproportionate with the flow velocity in a stream, it is argued to be important to study the gradient (Yochum & Bledsoe, 2010). As such, an increase of flow resistance could be the result of a decrease in maintenance condition of the stream. During this study, a new approach is be developed which is able to find the locations with decreased maintenance condition of the drainage area and to identify its causes. The approach uses machine learning techniques and anomaly detection on stream gradients on the main waterways of the drainage areas. This is studied because the slope of the stream gradient could be the result of the resistance in the stream.

Anomaly detection or outlier detection is a method that is specialized into finding patterns in data that are not in line with expectations considering the bulk of the data (Hodge & Austin, 2004). This method has several applications, such as fraud detection, medical applications, text errors and others. All of these applications have different anomaly detection techniques. There is an overview by Chandola et al. (2009) which reviews a number of these techniques and the advantages and disadvantages are given (Chandola, Banerjee, & Kumar, 2009). The most notable advantage of using the method for this study is that it is relatively simple. However, a disadvantage of using this method could be the amount of data needed, but because of a large available dataset this will not be an issue in this study.

For each problem where anomaly detection could be used, a decision has to be made on what specific technique is to be used. During the study, this decision is made considering the problem and the available data. Anomaly detection could be used for problems in water management as well. Raciti et al. (2012) describe the detection of contamination in water distribution systems. The ADWICE algorithm is used to find anomalies in the water quality of the system, real-time (Raciti, Cucurull, & Nadjm-Tehrani, 2012). In the Materials & Methods chapter of this proposal, a more in-depth description of the methods is given.

Most of the algorithms based on anomaly detection are classified as machine learning algorithms. Machine learning is divided into two main classes, supervised and unsupervised machine learning. Supervised machine learning is a technique where an algorithm is given a certain amount of training data in order to find relationships between the data. The training data is well monitored and the reasons for the anomalies are known (Kotsiantis, Zaharakis, & Pintelas, 2007). By analysing the training data, the algorithm produces a function that describes the data and is able to predict patterns. The perfect supervised learning algorithm will detect all the anomalies, labels them correctly and is able to predict anomalies.

Unsupervised machine learning does not use training data, but detects patterns based on clustering of the data. It leaves the steps of the algorithm unknown to the user (Kotsiantis, Zaharakis, & Pintelas, 2007). The latter method was not used in this study since the lack of data and the complexity of the method.

The stream gradient of a stream is the difference between the upstream and downstream water-level over the length of the stream. The steepness of the gradient is dependent on the flow velocity, the resistance of the stream and other factors (Ferguson, 2012). Overall, the flow velocity is relatively low in the study area, resulting in the fact that the relative dependency of the resistance to the gradient is high. An unusual high resistance in the stream could imply a bad maintenance condition. Since the machine learning domain is relatively new and changing fast, combining it with stream gradients in drainage areas is not yet present in the literature and will be done during this study. Since the stream gradient is proportional to the difference between the upstream and downstream water level (Δh), this value is used during the study in order to make it more comprehensible.

Anomaly detection and machine learning have a number of different techniques. The techniques used in this study are described in the next chapter of this report.

1.4 Research aim

The aim of the study is to develop an approach which is able to detect the maintenance condition of the main waterways of the drainage areas of HDSR which should result in better insights in the maintenance condition of drainage areas. This is important for increasing the efficiency of the maintaining schedule of the drainage areas of which HDSR in generally will benefit. In addition, the study will try to fill the gap in the literature on combining machine learning techniques with stream gradients. The following research question will be answered by the study.

To what extend is it possible to detect unwanted consequences of a decreased maintenance condition on the main waterways in drainage areas of HDSR, with the help of machine learning techniques?

In order to make the study and report clearer and more comprehensible, the main research question is divided into five sub-questions.

1. *What are unwanted consequences of a decreased maintenance condition of the main waterways in the drainage systems of HDSR?*
2. *Which machine learning technique is most suited to detect unwanted consequences of a decreased maintenance condition of the main waterways in draining areas of HDSR?*
3. *To what extend does the approach work for the Amerongerwetering drainage area specific?*
4. *To what extend does the approach work for the Lange Weide drainage area specific?*
5. *What kind of anomalies is the developed approach able to detect?*

2 Site description and theory

2.1 Site description

The area that is managed by HDSR reaches from Schoonhoven to Rhenen, bordered by the Nederrijn and the Lek. In the Netherlands, the main rivers are managed and measured by the Department of Waterways and Public Works. The other surface water bodies are managed and measured by the regional water authorities, such as HDSR. The first case study is conducted in the Amerongerwetering drainage area, which is located in the south-eastern part of the managed area of HDSR, as seen in figure 1.1. The second case study is conducted in the Lange Weide drainage area, which is located in the western part of the managed area of HDSR. Since the research consists of two case studies, these two drainage areas are the research areas.

2.1.1 The Amerongerwetering drainage area

The Amerongerwetering is perfect for the first case study as the input of water is provided from a single geographical location and the design of the system is rather simple, since there is a natural gradient present that controls the downstream discharge (Kort, 2010). The elevation of the drainage area ranges between +3 and +5 meter above NAP, the Dutch mean sea level (Actueel Hoogtebestand Nederland, 2020). The subsurface of the Amerongerwetering drainage area consists mostly of river clay above a layer of sand (TNO, Geologische Dienst Nederland, 2020). The clay and the sand are part of the Echteld Formation, deposited in the Holocene (TNO, Geologische Dienst Nederland, 2020).

In some locations there is a small layer of peat present between the clay and the sand. Figure 2.1 shows a schematic view of the subsurface at the location of the borehole. The clay is deposited by the flooding of the Nederrijn, just south of the drainage area. All of these boundary conditions make this first case study well constrained, so that the research is focussed on the methodological developments. As seen in figure 2.1, the Amerongerwetering drainage area is located relatively close to the Nederrijn river and the Kromme Rijn river. These two rivers could have an influence on the water level of the Amerongerwetering, which is considered in the research. Due to the fact that there are 3 measuring locations and weirs present in the Amerongerwetering, the waterway is divided into 2 parts. The upper part is located between the weirs Kolland and Nooit Gedacht and is about 700 meters in length. The lower part of the Amerongerwetering is located between the weir Nooit Gedacht and the location where it discharges into the Kromme Rijn. The length of the lower part is about 4.8 kilometres.

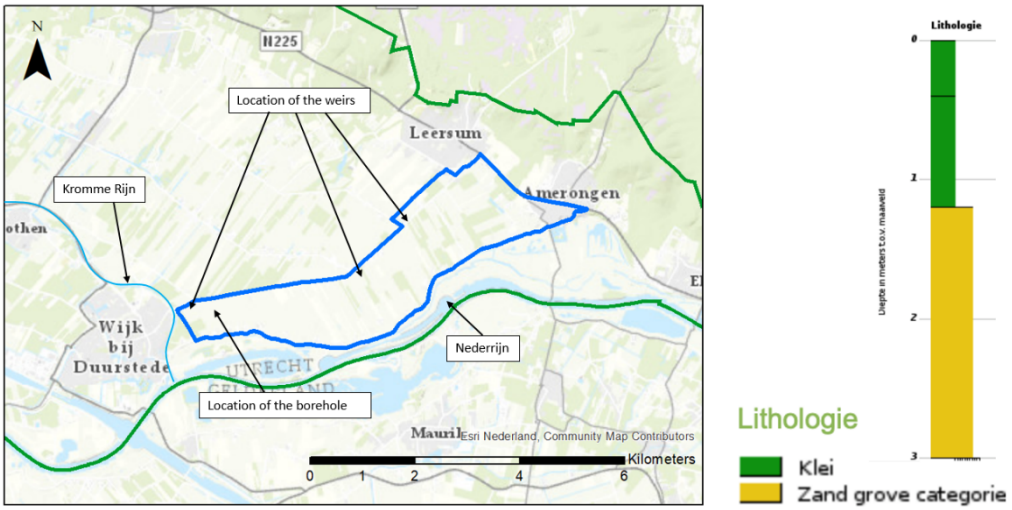

Figure 2.1: Location of the Amerongerwetering drainage area with the location of the borehole and the lithology of the subsurface (TNO, Geologische Dienst Nederland, 2020).

2.1.2 The Lange Weide drainage area

The second case study was performed in the Lange Weide drainage area, in the western part of the managed area of HDSR. The more complex situation of this drainage area makes it perfect for testing whether the approach is suitable for more complex situations as well. A long data record is available for the Lange Weide drainage area and several changes in the hydrology have been implemented over the years in order to decrease the land subsidence.

The subsurface of the Lange Weide drainage area consists of different layers of peat with smaller layers of clay in between, as seen in figure 2.2. The top clay layer of the schematic view of the subsurface, in figure 2.2, is part of the Echteld Formation. Below that, about 7 meters of peat is present, alternating with small clay layers, which is part of the Nieuwkoop Formation. This is deposited in the Holocene as well (TNO, Geologische Dienst Nederland, 2020). Below this layer of peat, sand is present of the Boxtel formation which is deposited during the previous glacial period, which ended about 11.65 thousand years ago, and after (Stouthamer, Cohen, & Hoek, 1996).

One of the first Dutch experimental "underwater drainage systems" is located in Lange Weide, designed to decrease land subsidence caused by peat oxidation, which is irreversible. This specific technique uses tile drains that are situated below the surface water level to decrease the range in groundwater level between periods. This technique assists to link the groundwater level to the surface water level, in a way that the groundwater level is managed more easily (Nationaal kennisprogramma bodemdaling, 2019).

Furthermore, there is a network of pressure drainage located near Lange Weide. This network could have an influence on the hydrological situation at the Lange Weide drainage area (Nationaal kennisprogramma bodemdaling, 2019). The elevation of the drainage area is around 2 meters below NAP, resulting in a much more complex situation without a strong natural gradient (Actueel Hoogtebestand Nederland, 2020). A pumping station is located in the drainage area which generates the gradient automatically.

As a result of the complexity of the Lange Weide drainage area, HDSR monitored it for a longer period of time and there is a maintenance dataset available. Figure 2.2 shows that the Lange Weide drainage area is located between the Oude Rijn river and the Hollandsche IJssel river. This is considered in the research, since it could have an influence on the water levels in the Lange Weide drainage area. Since there are no weirs situated in the main waterway of the Lange Weide drainage area, this area is not divided.

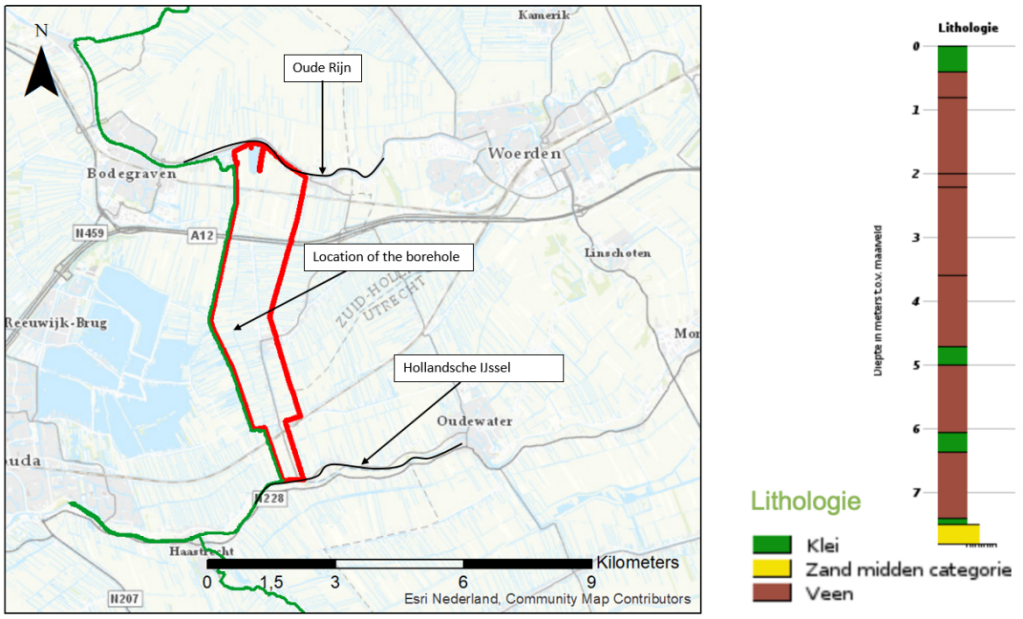

Figure 2.2: Location of the Lange Weide drainage area with the location of the borehole and the lithology of the subsurface (TNO, Geologische Dienst Nederland, 2020).

2.2 Machine learning techniques

2.2.1 Linear regression model

A linear regression model is a simple model that describes a situation with a linear regression equation (Zeileis, Leisch, Hornik, & Kleiber, 2001). Given below is an example of such an equation:

$$y = ax + b \tag{1}$$

$$y = a_1x_1 + a_2x_2 + a_3x_3 + b \tag{2}$$

Such a linear regression model is used in this study. The gradient of the stream is represented by the *y* variable. The *x* in equation 1 can be represented by different variables like, precipitation, evaporation, water-level. The linear regression could be extended to more than one variable, as the gradient is dependent on more variables. Such a model is called a multivariate linear regression model (Kabe, 1963) and an example is given in equation 2. The *a* and *b* in the equation are two constants, *a* is a specific constant linked to the variable *x* and *b* is a constant which is not linked to a variable in the equation. However, the situation might be too complex for such a linear regression model. So different models for the study are investigated as well, such as a gradient boosting model and a random forest model. Both of these models are predictive models based on decision trees.

2.2.2 Decision tree

A decision tree breaks down data in smaller sets with the help of decisions (Priyam, Abhijeet, Gupta, & Srivastava, 2013). Figure 2.3 is an example of a simple decision tree. The data is divided into four subsets with the help of 3 decisions. Decision tree could have an infinite number of layers and thus an infinite number of subsets. Both a gradient boosting model and a random forest model use decision trees in order to predict the relationship between variables.

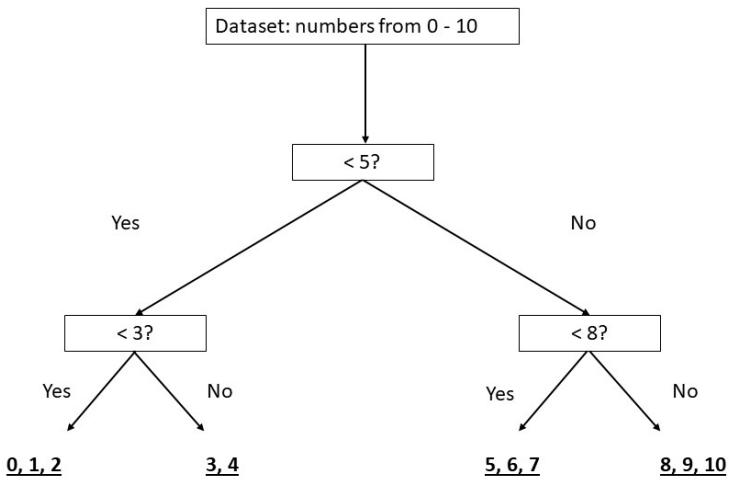

Figure 2.3: Example of a simple decision tree.

2.2.3 Gradient boosting model

Gradient boosting is a machine learning technique which produces a model out of several different decision trees in series. The input of the first decision tree is the training data. The output of tree number 1 is used as input for tree number 2 in order to learn from the process of tree number 1. This is commonly done by increasing the weights of the wrongly classified results, causing the second tree to focus more on classes that are harder to classify. This generally done with a number of different decision trees. An advantage of this method is that the processes of the different trees contribute to the total model. Many “weak learners” are converted into one stronger learner, which is called “boosting”, which helps decreasing the computing time (Natekin & Knoll, 2013).

2.2.4 Random forest model

Random forest is a technique which uses multiple decision trees in parallel, all using a small set of the training data. All the different trees run the process simultaneously and all have their own answer to the problem. The random forest method then reviews all the answers of the decision trees and decides which answer or solution to the problem is the most abundant in this list (Liaw & Wiener, 2002). The solution is chosen as the result of the random forest method.

3 Materials & Methods

3.1 General approach

The study consists of three phases, the introduction, data collection and data analysis phase. The introduction contains a thorough review of the previous studies on a similar subject, which is mostly described in the previous chapters. The data collection and data analysis phase are described in this chapter.

3.2 Data collection

3.2.1 Input datasets

There were three different kinds of datasets used during the research, which are listed in table 3.1. Firstly, the meteorological data used during the research was collected by the Royal Dutch Meteorology Institute, the KNMI, in De Bilt and in Cabauw (Koninklijk Nederlands Meteorologisch Instituut, 2020). This dataset originally consists of 39 different variables and was downloaded from the website of the KNMI. Before the dataset was used, the time components of some variables were deleted. Since these variables were not used in the research, 26 variables remained. The KNMI data has been made freely accessible, as it is part of the Ministry of Infrastructure and Water Management.

Secondly, the water levels in and around the drainage areas were included. The water levels in the drainage areas resulted in the difference in water level, Δh called in the rest of the report. The water levels outside of the drainage areas can have an influence on the Δh in the studied stream. For the Amerongerwetering drainage area this is the Nederrijn and the Kromme Rijn, for the Lange Weide drainage area this is the Oude Rijn and the Hollandsche IJssel. The data for this dataset is collected by HDSR and the Department of Waterways and Public Works.

The last dataset used for the research consists of area-averaged values of the precipitation and evaporation, collected by HDSR. These values contain the daily amount of precipitation and the actual evaporation. Using R programming language, a routine was developed to calculate the weekly, 2-week and 3-week mean values. These mean values were needed in the research in order to let the model to get familiar with long term patterns that are not visible in daily values. The three described datasets were combined into one dataframe with the help of a R script.

Table 3.1: List of all variables with their description and unit that are part of the input for the model.

Variable abbreviation	*Description*	*Unit*
DDVEC	Vector averaged wind direction	Degrees
FHVEC	Vector averaged wind speed	m/s
FG	24-hours averaged wind speed	m/s
FHX	Maximum hour-averaged wind speed	m/s
FHN	Minimum hour-averaged wind speed	m/s
FXX	Maximum wind gust	m/s
TG	24-hours average temperature	°C
TN	Minimum temperature	°C
TX	Maximum temperature	°C
T10N	Minimum temperature at 10 cm from surface	°C
SQ	Duration of sunshine (calculated from Q)	h
SP	Percentage of longest possible sunshine	-
Q	Global radiation	J/cm^2
DR	Duration of precipitation	h
RH	Daily precipitation	mm
RHX	Maximum hourly precipitation	mm
PG	24-hours averaged air pressure converted to mean sea level, calculated from hourly values	hPa
PX	Maximum hourly value of air pressure converted to mean sea level	hPa
PN	Minimum hourly value of air pressure converted to mean sea level	hPa
VVN	Minimum sight	m
VVX	Maximum sight	m
NG	24-hours averaged cloud cover	-
UG	24-hours averaged relative humidity	-
UX	Maximum relative humidity	-
UN	Minimum relative humidity	-
EV24	Reference crop evaporation	mm
PD	Area-averaged precipitation per day, a corrected value of the KNMI data for the research area.	mm
P1W	One week running mean value of PD	mm
P2W	Two week running mean value of PD	mm
P3W	Three week running mean value of PD	mm
ETD	Area averaged actual evapotranspiration per day	mm
ET1W	One week running mean value of ETD	mm
ET2W	Two week running mean value of ETD	mm
ET3W	Three week running mean value of ETD	mm
OR.sluice.down	Water level of the Nederrijn downstream of sluice Amerongen	m
OR.sluice.up	Water level of the Nederrijn upstream of sluice Amerongen	m
KR.AW	Water level of the Kromme Rijn at the location where the Amerongerwetering discharges in the Kromme Rijn	m
KR.MW	Water level of the Kromme Rijn, approximately 1 km downstream of discharge point of Amerongerwetering into Kromme Rijn	m
KR.WbD	Water level of the Kromme Rijn, approximately 1 km upstream of discharge point of Amerongerwetering into Kromme Rijn	m
OR.WbD	Water level of the Nederrijn at start of Kromme Rijn	m
Enkele.Wiericke	Water level of the Enkele Wiericke water way, the western border of the Lange Weide drainage area	m
Oude.Rijn	Water level of the Oude Rijn river, near the village of Bodegraven	m
Dubbele.Wiericke	Water level of the Dubbele Wiericke water way, the eastern border of the Lange Weide drainage area	m
Hollandsche.IJssel	Water level of the Hollandsche IJssel river, south of the Lange Weide drainage area	m
peil	Factor whether the water level is at summer or winter level	-
Δh	The difference between the upstream and downstream water level of a waterway	m
date	The date	-

3.2.2 Accuracy of the datasets

For this research, the KNMI meteorological data proved most accurate since the data was acquired directly from the source, which quality is controlled extensively by the KNMI. However, since the data was collected at the meteorological stations in De Bilt and Cabauw, the geographical location of the data is not the same as the research area. This difference might lead to minor spatial and temporal mismatches in data, such as precipitation data. HDSR estimates the accuracy of the water level data that 90% of the time, the error is less than 10 cm. This is generally due to the possibility that the staff gauge in the water can subside.

Over the last years, HDSR has made a lot of effort to increase the accuracy of the water level data. This is done by recalibrating the staff gauges, controlling the state of the staff gauges and logging all situations considering the staff gauges. By doing this, HDSR tends to increase the quality of the water level measurements.

The area-averaged precipitation data that was used during the research is a calculation of the KNMI radar data in order to make it representable for a specific research area. KNMI estimates the accuracy of the data at 96 – 98%. However, HDSR assumes that the accuracy of the area-averaged precipitation data is lower, since it is dependent of more factors than just the KNMI data. An estimation of the accuracy is not given by HDSR and was not calculated during this study since it is out of the scope of the study.

There is no estimation for the accuracy of the evapotranspiration present either, since this value depends on a number of different meteorological input factors. The data is achieved by eLEAF for HDSR; however, no accuracy is provided (eLEAF, 2020). It happens to be difficult to give a clear estimate of the accuracy, since a lot of different atmospheric processes have an influence on the evapotranspiration, as well as biological processes. Since it is out of the scope of this study, no estimate was calculated during this study.

3.2.3 Vegetation cover

HDSR measures the vegetation cover on several locations. This is done by logging the impression of the vegetation cover. This data is an impression and thus rather subjective and there are several data gaps present in the dataset as well. In 2019, consultancy firm Nelen & Schuurmans studied the vegetation cover measuring method. The firm concluded that the way of measuring has changed in January 2015. Before 2015, vegetation as algae, duckweed and other floating weed were included in the measurements. Since 2015, these forms of floating vegetation are not part of the measurements anymore, resulting in a shift in measurement output. The firm also concluded that the staff who performed the measurements changed over time, resulting in a shift in measurement results once more (Nelen & Schuurmans, 2019). The numerous data gaps in the vegetation cover dataset made it impossible to use it as an input for the models. However, the vegetation cover was used in the anomaly detection process of this research, which is described later in this chapter. Figure 3.1 shows the vegetation cover over time for the Amerongerwetering and the Lange Weide drainage areas.

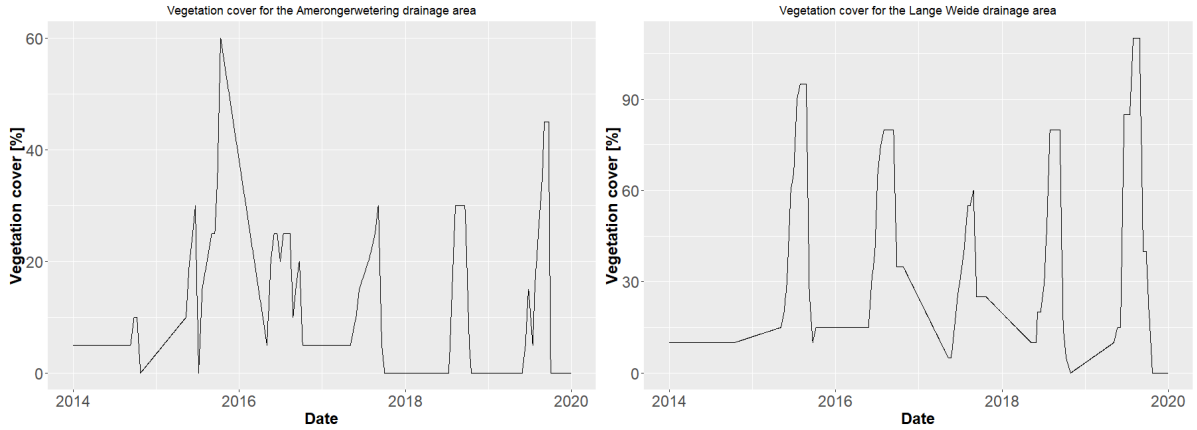

Figure 3.1: The vegetation cover over time for the Amerongerwetering (left) and Lange Weide (right) drainage areas.

3.3 Data analysis

In order to analyse the data for the research, different techniques were used. First a linear regression model was created in the software R. This linear regression model is a simplified version of the approach that was developed during this study. By studying this linear regression model, important lessons were learned which were used when developing the approach. As a result, the linear regression model acted as a benchmark for the other two models. Second, the random forest and gradient boosting models, which are described before, were used to analyse the data as well. These models were used to predict the Δh of the waterways assisted by all the input data.

3.3.1 Training and test dataset diversion

Before the models were able to predict Δh, training of the models had to be performed. In order to train the models, a training dataset is required. Training of the model is done by running it and let the model predict values, according to the input data. When the training is done, the model can be tested with the help of a test dataset. The testing is done in order to assess whether the predicted values of the model are similar to the observed values in the test dataset.

Although, there is no scientific consensus about a specific diversion ratio between the training dataset and the test dataset in the hydrology department, it is clear that the majority of the data should be used for training. In order to reduce overfitting, in this study a portion of ⅓ of the total dataset was used for testing (ResearchGate, 2020). The other ⅔ was used for training, since most data should be used for training. When using data that is dependent on time, for instance precipitation or water level, usually the first ⅔ part of the input dataset is taken as training dataset and the last ⅓ part is taken as test dataset. However, when this was done sequentially during the research, the results of the model were not accurate. The models had some problems with predicting longer periods with low precipitation rates and low water levels, as these periods were not present in the training dataset. A reason for this was that the summer of 2018, a relative dry period, was not part of the training dataset but was part of the test dataset. This resulted in a relatively low quality of the results during this period. This was a problem for the approach, since longer periods of dry conditions are prone to occur in the future. This problem was fixed by assigning the training dataset and test dataset at random points in time. As a result, the diversion ratio did not change. However, the two datasets did change.

3.3.2 Settings of the models

Both the random forest and the gradient boosting model are found in the R package “caret”. This free package consists of a set of functions that assist in creating predictive models. Prior to the running, the models require specific settings in order to prepare them for the data. For both the random forest as the gradient boosting model the training dataset was divided into 10 folds cross-validation (*number*). The *repeat* was set to 3, which means that the model is required to complete 3 sets of folds (Kuhn, 2020).

The random forest model only needs the *mtry* set to the situation. This setting is the number of randomly collected variables to be sampled at each split time. In an ideal situation this number would be the total number of variables. However, since the model input consists of 38 to 43 variables, depending on the model, the computing time would exceed the computation resources in this study. The *mtry* was optimized by letting the model run for different *mtry* settings. The *mtry* was set as 24, since this was an optimal value compared to the computing time and the quality of the model.

The gradient boosting model has several other settings, such as *n.trees*, *interaction.depth*, *shrinkage* and n.*minobsinnode* (Kuhn, 2020). The values of these settings were obtained with a similar method as the *mtry*. The computing time and quality of the model compared for different settings. The *n.trees* setting represents the number of trees or iterations the model used in the computation. This value was set to 3000.

The *interaction.depth* stands for the complexity of the tree (i.e. the numbers of splits of a single decision tree). This value was set to 5, meaning that the total number of terminal nodes per tree is 6.

Shrinkage represents the learning rate of the model meaning that this factor signals the model when to stop. The used value of 0.01 for this study signals the model to stop when the difference between the different trees is less than 0.01.

The final setting is called the *n.minobsinnode*, which represents the minimal number of observations per node. When a tree ends with less observations in a node, the model is stopped as well. For this study, this value was set to 5. The previously described settings and their values are shown in table 3.2.

Table 3.2: Settings of the models and their explanation and value (Kuhn, 2020).

Setting	*Explanation*	*Value*
number	Number of fold cross-validation	10
repeats	Required number of complete folds	3
mtry	Number of randomly collected variables to be sampled at each time split	24
n.trees	Number of trees	3000
interaction.depth	Number of splits per tree	5
shrinkage	Learning rate	0.01
n.minobsinnode	Minimal number of observations per node of a tree	5

3.3.3 Anomaly detection

The number of predicted values differs for each area, since *NA* value datapoints result in model errors. Therefore, the decision was made to delete all dates with a *NA* value in one or more of the variables from the input. Since each area has different input data, the total number of train observations differs for each location. This resulted into a different number of predicted values per location. Since the top rank and bottom rank 5% of the residual values are classified as anomalies, 50% of the anomalies were predicted with a too large value and the rest with a too low value of Δh.

After running the models, the predicted Δh was compared with the observed Δh from the test dataset. This was done with equation 3, which calculates the difference between those values. In equation 3, Δh_{obs} is the value of the observed and Δh_{pred} the value of the predicted Δh. The residual value was used to find the date which are not predicted well enough.

For this research, the top rank 5% and the bottom rank 5% of the residuals were classified as "not sufficient predicted", which are called the anomalies. The reason for classifying the top rank and bottom rank 5% of the residual factors as anomalies is because these values are not in order with the rest of the data. These top rank and bottom rank 5% represent the anomalies that are studied in this research. Classifying anomalies with the help of the top rank and bottom rank 5% helps find anomalies in the positive and negative side of the histogram. Using the standard deviation when classifying the anomalies, several important anomalies would be lost due to the fact that the standard deviation focusses on two equal borders, positive and negative. Using 5% as the border for which the anomalies are detected, the border is area-specific. Since the water levels, and thus Δh, in both case study areas are different, no strict border can be taken. Creating a border which is based on a percentage overcomes this problem.

$$Residual = \Delta h_{obs} - \Delta h_{pred} \quad (3)$$

After the anomalies are detected, it is important to find possible reasons that explain why the model did not predict Δh sufficiently in such a situation. This process was done with the help of the vegetation cover data and the water level data of the researched waterways in the Amerongerwetering and Lange Weide drainage areas. As is previously described, the vegetation cover is a rather subjective dataset since it is an impression and there are several data gaps present. Therefore, the vegetation cover was not used as an input for the model. However, the quality of the data is appeared to be sufficient enough to visually compare the anomalies with this dataset. The vegetation cover patterns were studied over a period of a few days and the relative change was used

to compare the anomalies. Examples of vegetation cover patterns are given in table 3.3 including the predicted effect on Δh. The water levels in the drainage areas were used to find explanations for the anomalies as well. The patterns in the water levels were added in table 3.3 in addition, with the expected effects.

Table 3.3: Vegetation and water level pattern with their possible explanation and predicted effect on Δh.

Vegetation cover pattern	**Possible explanation**	**Predicted effect on Δh**
Sudden steep decrease	Mowing of the vegetation	Decrease of Δh afterwards
Increase over time	Growing of the vegetation	Slow increase of Δh during this period
No change, after a period of increase	End of growing season	After a slow increase, Δh now does not change
Water level pattern	**Possible explanation**	**Predicted effect on Δh**
Sudden increase difference between the water levels	A weir has been closed more than usual	Increase
Sudden decrease in difference between the water levels	A weir has been opened more than usual	Decrease
A negative difference between the water levels	A peak in precipitation	Negative value

To decide which model suits each drainage area, the *RMSE* value of the models was used. This value is calculated with equation 4 and gives the root-mean-squared error of the model. This equation has proven to be useful to compare the different models for the same data; a comparison within the case studies. For comparison of the quality of the models outside the case studies, the R^2 value is used. This value provides a better view of the quality of the models with different input data and their temporal dynamics. The *RMSE* value ranges from 0 to 1, the lower the *RMSE* value, the better the fit of the predicted values to the observed data. The R^2 value ranges from 0 to 1 as well. However, the an R^2 value of 0 implies no correlation between the input data and the observed Δh.

$$RMSE = \sqrt{E((\Delta h_{obs} - \Delta h_{pred})^2)} \tag{4}$$

3.3.4 Unwanted consequences of decreased maintenance condition

In order to study the maintenance condition of drainage areas, it is needed to find the unwanted consequences of such. The data that is available for the drainage areas (i.e. the input data and the logging data of situations at the weirs) was studied in order to find these unwanted consequences. This is described in the introduction phase of the research, since the outcome of this is needed during the research phase. By studying the data, there was found that the most important factor is the fact that non-mowed vegetation decreases the maintenance condition of the drainage area. An excessive amount of vegetation in the water and on the banks could lead to increased flow resistance, which is classified as a decrease in maintenance condition since it could increase flood risk (Darby, 1999).

A broken weir fails to do the job it is built for, which could decrease an influence on the maintenance condition. However, such a situation has not occurred in the study period at any of the studied weirs. Other factors like trees that might have fallen into the water, broke piping under roads, broken pumping stations have similar effects. However, these situations were not present in the study period. The total influence of these consequences is considered small because of a significantly low probability of occurrence.

4 Results

In this chapter, the results of the linear regression model are described, followed by the results of the two case studies. The Amerongerwetering was divided into an upper and lower part, since a weir is located in the drainage area. For each case study, the results of the random forest model and the gradient boosting model will be described subsequently. As explained in the previous chapter, for each model, the top rank 5% and bottom rank 5% of the residual values per model are classified as anomalies.

4.1 Linear regression model

The linear regression model was developed in the preliminary part of the study. After running the model, the *RMSE* values revealed that the research problem was too complex for a linear regression model to solve. The model that consisted of the input data for the lower part of the Amerongerwetering resulted in a *RMSE* value of 0.145. Due to these results, the linear regression model was removed from the study. Although the R^2 value could provide insights to compare different areas, this value was not calculated, since a linear regression model was only applied to the lower part of the Amerongerwetering.

4.2 Results of the case studies

The number of predicted values of Δh differs per model, due to the fact that the dates with *NA* values for one or more variables were deleted from the input dataset. Since this is different for the separate areas, the total number of predicted values differ per area. Ten percent of the predicted values were classified as anomalies, so the number of anomalies differ per area as well. These values are shown in table 4.1. The number of anomalies were equal for the different models of the same area. However, these anomalies could represent different dates and values. The *RMSE* and R^2 values are shown in the table as well.

Table 4.1: Number of predicted values, anomalies and the RMSE and R^2 values per model.

Model and area	*Number of predicted values*	*Number of anomalies*	*RMSE*	*R^2*
Random forest for upper Amerongerwetering	696	70	0.0829	0.6719
Gradient boosting for upper Amerongerwetering	696	70	0.0814	0.6718
Random forest for lower Amerongerwetering	571	58	0.0778	0.7310
Gradient boosting for lower Amerongerwetering	571	58	0.0849	0.6690
Random forest for Lange Weide	587	60	0.0592	0.2977
Gradient boosting for Lange Weide	587	60	0.0550	0.3737

In order to statistically compare the results of the different models, the *RMSE* value was used. The *RMSE* values in table 4.1 show that for the upper part of the Amerongerwetering the gradient boosting model is better. However, the difference between the values is not statistically significant. For the lower part of the Amerongerwetering, the random forest model scores better. The models for the Lange Weide drainage area have lower but similar *RMSE* values, which were lower than the values of the other areas. It was impossible to distinguish the better model based on these values.

The models show rather different results, as is shown on the following pages. The R^2 was used to compare the results of the different areas. Table 4.1 shows that, for the upper and lower part of the Amerongerwetering, the models display relatively high R^2 values as compared to the results of the Lange Weide models. This implies that the models are less suited to predict the Δh for the Lange Weide drainage area.

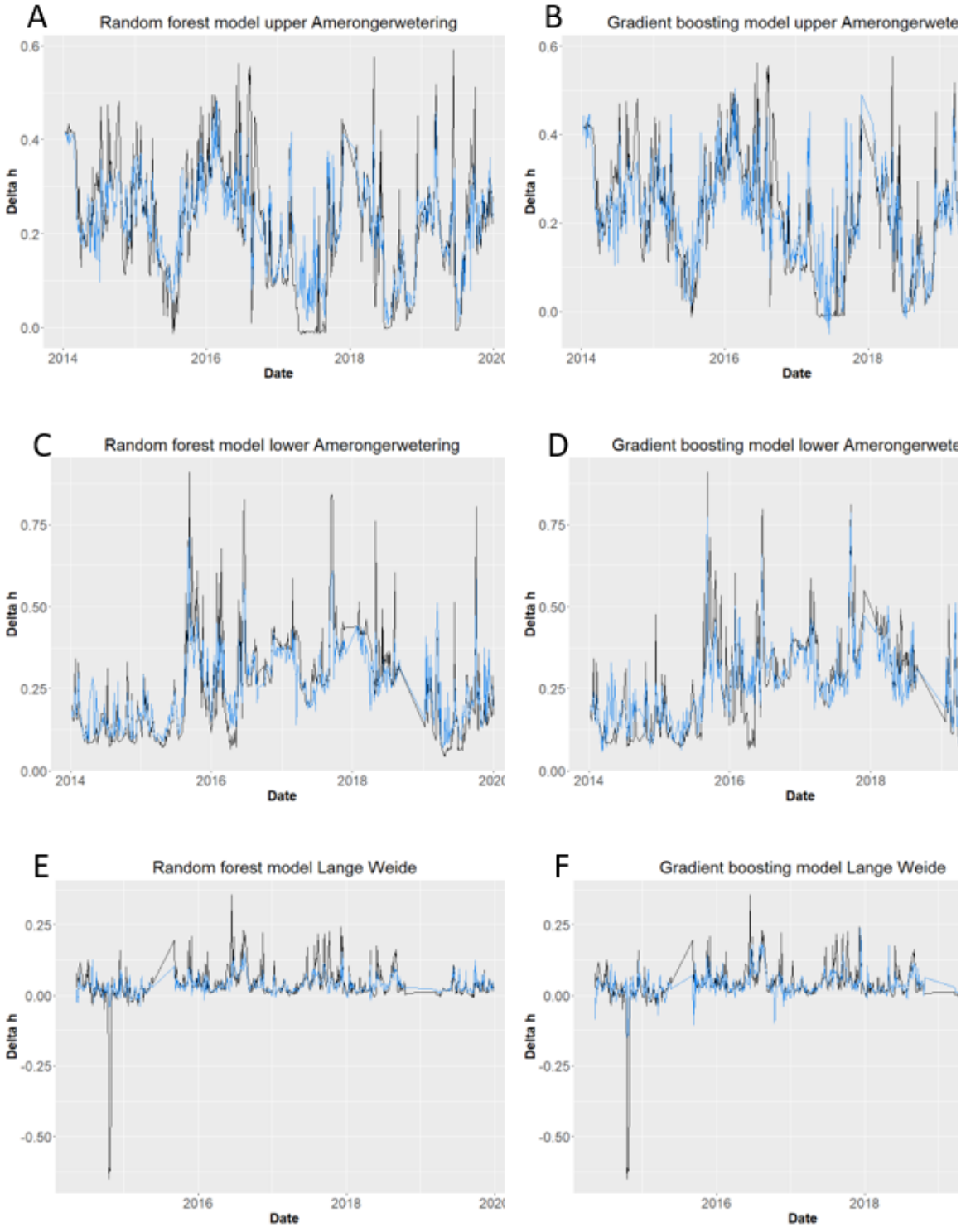

Figure 4.1: The observed (black) and predicted (blue) difference in water level. A: Random forest, upper Amerongerwetering; B: Gradient boosting, upper Amerongerwetering; C: Random forest, lower Amerongerwetering; D: Gradient boosting, lower Amerongerwetering; E: Random forest, Lange Weide; F: Gradient boosting, Lange Weide.

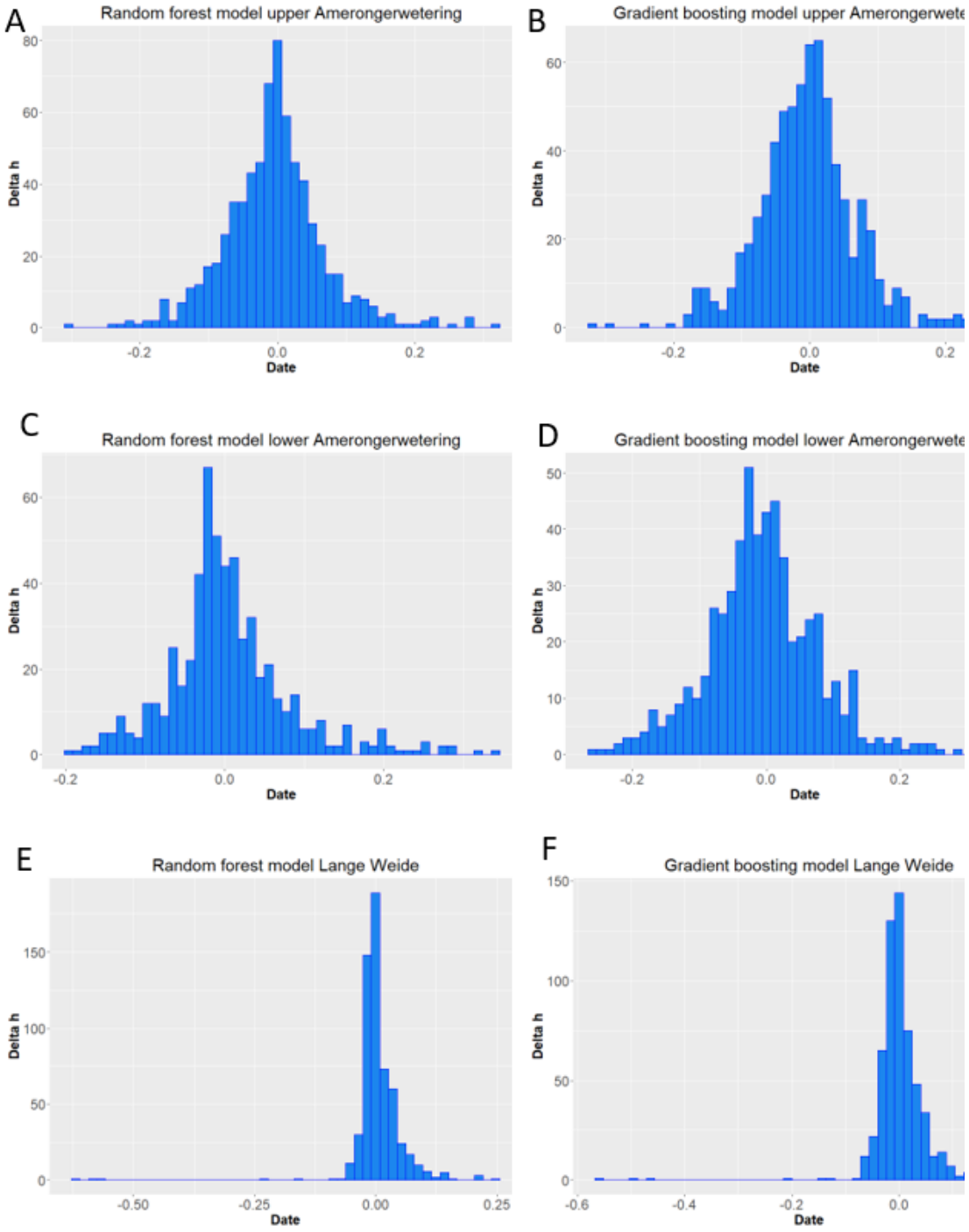

Figure 4.2: Histograms of the distribution of the residual values. A: Random forest, upper Amerongerwetering; B: Gradient boosting, upper Amerongerwetering; C: Random forest, lower Amerongerwetering; D: Gradient boosting, lower Amerongerwetering; E: Random forest, Lange Weide; F: Gradient boosting, Lange Weide.

Figure 4.3: The anomalies plotted over time with their possible explanation, based on appendices 1 -6. A: Random forest, upper Amerongerwetering; B: Gradient boosting, upper Amerongerwetering; C: Random forest, lower Amerongerwetering; D: Gradient boosting, lower Amerongerwetering; E: Random forest, Lange Weide; F: Gradient boosting, Lange Weide.

4.2.1 The upper part of the Amerongerwetering drainage area

The graphs clearly show that the results of the two models used for the same area, are rather different. In figure 4.1, it is clear that, for the upper part of the Amerongerwetering, the curve of the random forest model is a better predictor of Δh. Although the overall curve of the observed Δh is better predicted by the random forest model, the peaks in the observed curve do not always align with the peaks in the predicted curve. The gradient boosting model predicted these peaks better but showed more errors overall. These peaks are important, because of the fact that if the Δh at these dates is not predicted accurately enough, it implies that such conditions have not occurred during the training data. Since the training dataset is quite large, this means that these dates have specific conditions that do not occur often. By studying these conditions, the anomalies could be explained. The histograms in figure 4.2 shows that the residual values are distributed according to a normal distribution. It is clear that the distribution of the random forest model is better, since is it narrower to the top and the slopes of the histogram are smoother. This is another advantage of choosing this model over the gradient boosting model.

Figure 4.3 shows the spread of the anomalies over time with possible explanations on the y-axis. The colour of the dot indicates whether the anomalies are located in the top (red) or bottom (blue) rank of the residual values. The anomalies are evenly distributed over time. It should be noted that when an anomaly has multiple possible explanations, the dot in the figure is located in the row of the vegetation change. This way the number of anomalies that could be explained by the water level pattern seems lower than it really is. However, the number of anomalies with multiple possible explanation is low, so the effect of this is relatively small. The figures show that most of the anomalies

are classified under the “other” category. This is due to the fact that it is not always clear what caused the anomaly. An anomaly could be caused by one larger factor or by several smaller factors. The cause of an anomaly that is created by one large factor is distinguished more easily. The anomalies that are classified in the top three rows in the figures are anomalies that have at least one large factor of influence that is linked to the anomaly. Of these anomalies, vegetation change, both positive and negative, are possible explanation to a higher number of anomalies than the water level pattern in the upper part of the Amerongerwetering. This is due to the fact that this drainage area has a natural gradient, meaning no pumping station is needed to create a gradient. This natural gradient creates a stable water level pattern, resulting in a stronger influence of the vegetation cover on Δh

4.2.2 The lower part of the Amerongerwetering drainage area

In figure 4.1, the results of the lower part of the Amerongerwetering are shown. It is clear that the curve of the random forest model predicted the observed values better than the gradient boosting model. Similar to the upper part of the Amerongerwetering, the random forest model is clearly better in predicting the Δh values for this part of the drainage area. The histograms in figure 4.2 confirm this, as the histogram of the gradient boosting model is rather wide and shows rougher slopes.

The anomalies plotted over time (figure 4.3), show a pattern similar to the upper part of the drainage area. The anomalies are evenly distributed over time, except for 2018 and 2019 in the gradient boosting model. This period has a relative low number of anomalies. Most of the anomalies are classified in the “other” category, and for the rest of the anomalies, the vegetation is the most important possible explanation. Figure 4.3C shows that, in the random forest model, no anomaly can be explained by the water level pattern. However, eight anomalies not only had a water level pattern as a possible explanation, but a change in vegetation cover as well. Since the vegetation cover is thought to have a larger influence on the Δh, these anomalies are present in the vegetation change row of the figure.

4.2.3 The Lange Weide drainage area

The figures that show the results of the models on the Lange Weide drainage area are rather different from the other figures, since the Δh is relatively small compared to the Amerongerwetering drainage area. There is a large negative peak visible in figure 4.1, which should be interpreted as a measuring error. In a regulated polder, it is highly unlikely that the difference in upstream and downstream water level will decrease with about 55 centimetres in such a short period of time. Seeing is that both models do no predict this peak fully, it implies an extraordinary situation. While the curves of the Lange Weide models are rather different from the other models, a similar pattern is present, when looking at Δh. Compared to the gradient boosting model, the random forest model is clearly better in predicting Δh. The gradient boosting model has more errors and sometimes positive peaks are predicted as negative peaks. Although, the histograms of both models are rather similar (figure 4.2), the histogram of the gradient boosting model is wider, which is seen in the other areas as well. It can be concluded from the results that the random forest model is best in predicting the Δh for the Lange Weide drainage area.

Figure 4.3, that show the distribution of the anomalies over time, reveals that the anomalies are evenly distributed over the study period. A similar pattern is visible compared to the other areas, meaning that most of the anomalies are classified in the “other” category regarding possible explanations. The reasons for this classification are stated in section 4.2.1. The figures show that the water level pattern has a greater influence compared to the other two areas, as more anomalies can be explained by water level pattern. This is due to the pumping station creating water level patterns that are not always natural, such as a reversed gradient. The vegetation change has a similar number of anomalies compared to the other areas.

5 Discussion

It is important to assess the reliability of the results. In this chapter, the assumptions and choices made during the study are discussed and justified. Furthermore, the possible effects of these assumptions and choices are described. Next, it is discussed whether the results are reliable enough to answer the research questions. Lastly, recommendations are given on how to increase the reliability of the results and how to further study this topic.

5.1 Machine learning in literature

It is in the nature of the gradient boosting model to focus on peaks and outliers (Son, Jung, Park, & Han, 2015). Focussing on the peaks, logically results in a decreased focus of the model on the bulk of the data. During the study, this resulted in the gradient boosting models predicting the overall peaks better than the random forest models. However, the results also show that the random forest models predicted the non-anomalous values of Δh better than the gradient boosting models. This can be attributed to the random forest model does not focussing on the peaks, but having the focus divided randomly over the dataset (Biau, 2012). For this study, an increased focus on the peaks was not necessary.

Most of the studies using anomaly detection technologies in combination with water management focus on water quality, since it is an important part of water management and anomaly detection is of a great value for this topic (Raciti, Cucurull, & Nadjm-Tehrani, 2012), (Leigh, et al., 2019).

5.2 Assumptions

The first assumption underlying this study, is that the data, either used as input or during the analysis, is of sufficient accuracy and reliability. Although, measuring errors could be present in the data, most errors will have no significant influence on the results. The period 18/10/2014 to 31/10/2014 shows patterns in the observed values of Δh which can only be explained as a period of measuring errors at the inlet of the drainage area. The Δh value dropped from around 0 to -0.60 m in less than a day, resulting in a strong gradient towards the inlet of the drainage area, away from the pumping station. No logging data from the pumping station has been recorded for that period, implying that no abnormal events occurred. When combined with the fact that the drainage area is situated 2 meters below mean sea level, it is reasonable to conclude that an error occurred during these measurements. During the data processing, datapoints with any *NA* values or values that imply errors (e.g. -999.999) were deleted from the dataset, resulting in these datapoints being excluded from the analysis. Further periods of measurement errors were not detected during this study.

The vegetation cover data was not used as an input for the models, since the data was rather subjective and several data gaps were present. However, the data was used in the analyses. Originally it was planned to use a soil moisture dataset in the input for the models as well. However, the dataset suffered from multiple large data gaps, resulting in insufficient accuracy.

Another assumption in this study, was that all of the important influences on the water levels in the water ways were included. However, there might be influences that were not considered. For example, the fact that the grasslands next to the water ways are irrigated with water from the water ways. This was discovered when the report was being written. The amount of irrigation water that is extracted from the water ways in the drainage areas of HDSR is not measured or logged. Resulting in no available data for this study. However, future studies could include the factor by assuring that the amount of irrigation water is measured. This would increase the accuracy of the models and predictions, so it is recommended to include it in further research.

During the study, the value of Δh was used instead of the gradient. Since the gradient is proportional to Δh and this value is easier to interpret, the choice was made to use Δh.

Models are a simplified version of reality, and thus assumptions have to be made in order to make the models computable and comprehensible. During studies, it has to be clear that models per definition are not equal to reality.

A basic assumption during anomaly detection studies is that anomalies are the product of an unusual situation. This assumption is made in the early stages of the anomaly detection techniques and is copied for most of the anomaly detection studies (Gates & Taylor, 2006). This assumption could not be true for this specific topic, since it is not investigated. However, no indication is present that the assumption is not true, so the assumption is made for this study.

Another basic assumption for anomaly detection is that anomalies are assumed to be rare compared with normal data points (Chandola, Banerjee, & Kumar, 2009). While this assumption is generally true, it is not true that anomalies are always rare. Although anomalies are rare for this specific study, the measurement errors at the inlet of the Lange Weide drainage area at the period 18/10/2014 to 31/10/2014 show that this assumption has to be considered for each study separately. Since supervised anomaly detection models are used during this study, the errors in October 2014 were quickly discovered.

5.3 Recommendations

HDSR could use the results and outcome of the study to develop the tool further and use it for other drainage areas as well. By doing this, the condition of the waterways, and thus the drainage areas, could be assessed remotely. The data from this study is publicly available, resulting in no additional costs for data collection. Developing the tool into a working approach for all the drainage areas requires a rather simple follow-up study which could be performed in a relatively short period of time. However, some drainage areas might require more research due to the complexity of the hydrology of these areas. It is recommended that HDSR uses the results and outcomes of this study to further research and develop an approach for all the drainage areas. Studying the influential factors for each drainage area is important as well.

As stated in one of the previous sections, it is recommended that the effect of irrigation water being pumped from the water ways on the Δh is studied and considered during further research. Depending on the result of such a study, this factor should be either included or excluded in the approach. Since the effect was unknown in the current study, anomalies might be explained by the extraction of irrigation water by farmers.

It is recommended that further research redefines the boundary of the explanation of anomalies. Since the current boundary could be stricter, it should be studied how the approach defines and handles anomalies with multiple explanations. By doing this, more anomalies will be explained and the quality of the approach will increase.

6 Conclusion

The study concludes that unwanted consequences of a decreased maintenance condition of the main waterways in the drainage areas is an excessive amount of vegetation in and around the waterways. As a result of vegetation, flow resistance increases, which is considered unwanted since it could increase flood risk. Other factors as broken weirs, broken pumping stations, fallen trees into the waterways, and broken piping under roads are unwanted consequences as well. However, their influence is considered to be small.

The results of the study uncover that when the random forest model is used, the unwanted consequences are better detected. This can be explained by the fact that the prediction of the Δh is more accurate in the random forest models compared to the gradient boosting models. A gradient boosting model focusses not the peaks, while a random forest model focusses on all the data, which results in a better prediction by the random forest models.

As the *RMSE* values confirm, the approach is working well for the Amerongerwetering. This is clearly visible in the results as well. Further research can enhance the accuracy by making some adjustments, such as adding more data and studying the influential factors in the area. Further research can optimize the approach for the Lange Weide drainage area as well. This can be done by studying the influential factors in the drainage area and the complex situation more in-depth.

As seen in the results of the models, it is difficult to find explanations for the anomalies since most anomalies can occur because of multiple reasons. However, the approach was able to detect most anomalies. The process of explaining the anomalies takes time and further research to enhance. During this study, the detected anomalies were presumably caused by excessive amounts of vegetation, a drop of amount of vegetation cover and some extraordinary water levels.

Finally, the study concludes that machine learning provides opportunities for the water management in the drainage areas of HDSR. However, the results differ between the different study areas. The approach works relatively well for the Amerongerwetering drainage area due to its simplicity, compared to the Lange Weide drainage area. For this area, the approach does not work well, due to the pumping station and the complexity of the drainage area. This implies that the approach works well for relatively simple drainage areas, compared to areas that are considered complex. Most of the anomalies were detected by the approach. I recommend to further study the opportunities that machine learning and anomaly detection have for water management of the drainage areas.

References

Yochum, S., & Bledsoe, B. (2010). Flow resistance estimation in high-gradient streams. *2nd Federal Interagency Conference, Law Vegas, NV, June*.

Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. *Atrificial intelligence review*, 85-126.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys*, 1-58.

Raciti, M., Cucurull, J., & Nadjm-Tehrani, S. (2012). Anomaly detection in water management systems. *Critical infrastructure protection*, 98-119.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 3-24.

Ferguson, R. I. (2012). River channel slope, flow resistance, and gravel entrainment thresholds. *Water Resources Research*.

Kort, J. W. (2010). *De Amerongerwetering: een detailstudie met SOBEK channel flow (Doctoral dissertation).* Utrecht.

Actueel Hoogtebestand Nederland. (2020, February 17). *Actueel Hoogtebestand Nederland viewer*. Retrieved from website of Actueel Hoogtebestand Nederland: https://www.ahn.nl/ahn-viewer

Nationaal kennisprogramma bodemdaling. (2019). *Factsheet onderwater- en drukdrainage.*

Zeileis, A., Leisch, F., Hornik, K., & Kleiber, C. (2001). strucchange: An R package for testing for structural change in linear regression models. *Journal of statistical software*, 1-38.

Priyam, A., Abhijeet, Gupta, R., & Srivastava, S. (2013). Comparative analysis of decision tree classification algorithms. *International Journal of current engineering and technology*, 334-337.

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, 21.

Liaw, A., & Wiener, M. (2002). Classification and Regression of randomForest. *R news 2*, 18-22.

TNO, Geologische Dienst Nederland. (2020, April 20). *Ondergrondgegevens*. Retrieved from Dinoloket: https://www.dinoloket.nl/ondergrondgegevens

Stouthamer, E., Cohen, K., & Hoek, W. (1996). De vroming van het land. In E. Stouthamer, K. Cohen, & W. Hoek, *De vorming vna het land* (p. 248). Utrecht: Perspectief Uitgevers.

Kabe, D. (1963). Stepwise Multivariate Linear Regression. *Journal of the American Statistical Assiciation*, 770-773.

Koninklijk Nederlands Meteorologisch Instituut. (2020, March 5). *Klimatologie*. Retrieved from KNMI: http://projects.knmi.nl/klimatologie/daggegevens/selectie.cgi

eLEAF. (2020, May 16). *ET data*. Retrieved from eLEAF: https://eleaf.com/?page_id=3337

Kuhn, M. (2020, March 20). *The caret package*. Retrieved from Github: https://topepo.github.io/caret/index.html

Nelen & Schuurmans. (2019). *Analyse begroeiingsgraad.* Utrecht: Nelen & Schuurmans.

ResearchGate. (2020, May 29). *ResearchGate Questions*. Retrieved from ResearchGate: https://www.researchgate.net/post/is_there_a_formal_method_to_split_data_set_into_training_and_testing

Son, J., Jung, I., Park, K., & Han, B. (2015). Tracking-by-segmentation with online gradient boosting decision tree. *The IEEE International Conference on Computer Vision*, 3056-3064.

Darby, S. (1999). Effect of riparian vegetation on flow resistance and flood potential. *Journal of hydraulic engineering*, 443-454.

Biau, G. (2012). Analysis of a random forest model. *Journal of Machine Learning Research*, 1063-1095.

Gates, C., & Taylor, C. (2006). Challenging the Anomaly Detection Paradigm A provocative discussion. *NSPW '06: Proceedings of the 2006 workshop on New security paradigms* (pp. 21-29). Germany: Assiciation of Computing Machinery.

Leigh, C., Alsibai, O., Hyndman, R., Kandanaarachchi, S., King, O., McGree, J., . . . Peterson, E. (2019). A framework for automated anomaly detection in high frequence water-quality data from in situ sensors. *Science of The Total Environment*, 885-898.

Acknowledgements

I would like to thank Hoogheemraadschap De Stichtse Rijnlanden for making this internship possible. The organisation provided most of the data and a laptop to conduct the research on and all my direct colleagues. In particular I would like to thank my supervisor T. De Lange for his general help, sharing his knowledge and experience during the study and his feedback on the thesis.

I would like to thank dr. ir. N. Wanders for his detailed feedback and wish him, as well as the second reader dr. M. Van Vliet, well in reading the thesis and providing a mark.

I would like to thank my beloved friends L. Nibbeling, L. Doorakkers and father E. Roos for reading my thesis and giving clear feedback on how to improve it.

Appendices

Abbreviations in the appendices:
K = Water level at weir Kolland
NG = Water level at weir Nooit Gedacht
AW = Water level at location where Amerongerwetering discharges into Kromme Rijn
IN = Water level at inlet Lange Weide drainage area
PS = Water level at pumping station Lange Weide drainage area
Residual factor 1 = top rank 5% of the residual values
Residual factor -1 = bottom rank 5% of the residual values

Appendix 1: The anomalies of the random forest model for the upper Amerongerwetering

date	residual factor	current vegetation state	next vegetation state	relative change in vegetation state	interpretation vegetation	particularities water level Amerongerwetering	observed gradient of the Amerongerwetering	modelled gradient of the Amerongerwetering	residual
12/03/2014	1	normal increase	normal increase	/	non-growing season	NG increase	0,362	0,241	0,121
10/06/2014	1	relatively steep increase	relatively steep increase	/	growing season	/	0,281	0,150	0,131
16/06/2014	1	relatively steep increase	no change	negative	end of growing season	/	0,302	0,144	0,158
03/07/2014	1	no change	no change	/	growing season	/	0,292	0,156	0,136
05/07/2014	1	no change	no change	/	growing season	K increase	0,287	0,129	0,158
06/07/2014	1	no change	no change	/	growing season	K increase	0,289	0,097	0,192
19/08/2014	1	data gap	data gap	/	growing season	K increase and NG decrease	0,474	0,305	0,169
29/09/2014	1	normal increase	no change	negative	end of growing season	/	0,397	0,257	0,140
11/10/2014	1	no change	normal decrease	negative	mowing	/	0,481	0,333	0,148
19/12/2014	-1	steadily increase	steadily increase	/	growing season	K decrease and NG increase	0,136	0,308	-0,172
28/05/2015	-1	normal increase	normal increase	/	growing season	K decrease	0,026	0,152	-0,126

29/01/2016	1	possible data gap	possible data gap	/	non-growing season	K increase and NG decrease	0,453	0,329	0,124
01/06/2016	1	normal increase	no change	negative	growing season	K increase	0,516	0,339	0,177
14/06/2016	1	no change	normal decrease	negative	mowing	K increase	0,562	0,336	0,226
25/06/2016	-1	normal decrease	normal increase	positive	growing season	K decrease and NG decrease	0,264	0,390	-0,126
30/07/2016	1	no change	no change	/	growing season	K increase and NG increase	0,368	0,244	0,124
04/08/2016	1	no change	no change	/	growing season	peak K	0,550	0,404	0,146
10/08/2016	1	no change	steep decrease	negative	mowing	/	0,518	0,242	0,276
11/08/2016	1	no change	steep decrease	negative	mowing	/	0,555	0,279	0,276
15/08/2016	1	no change	steep decrease	negative	mowing	K decrease and NG decrease	0,367	0,239	0,128
25/08/2016	1	steep decrease	normal increase	positive	growing after mowing	/	0,282	0,158	0,124
26/08/2016	1	steep decrease	normal increase	positive	growing after mowing	NG decrease	0,288	0,091	0,197
02/09/2016	1	normal increase	normal increase	/	growing season	NG decrease	0,448	0,287	0,161
03/09/2016	1	normal increase	normal increase	/	growing season	both curves flat	0,448	0,237	0,211
04/09/2016	1	normal increase	normal increase	/	growing season	/	0,448	0,233	0,215
24/11/2016	-1	no change	no change	/	non-growing season	/	0,129	0,288	-0,159
26/11/2016	-1	no change	no change	/	non-growing season	/	0,104	0,241	-0,137
26/02/2017	-1	no change	no change	/	non-growing season	just after peak of K and NG	0,184	0,324	-0,140
27/02/2017	-1	no change	no change	/	non-growing season	/	0,168	0,328	-0,160
04/03/2017	-1	no change	no change	/	non-growing season	/	0,179	0,348	-0,169
06/03/2017	-1	no change	no change	/	non-growing season	/	0,210	0,368	-0,158
09/03/2017	-1	no change	no change	/	non-growing season	small peak K and NG	0,279	0,416	-0,137
17/03/2017	-1	no change	no change	/	non-growing season	/	0,090	0,219	-0,129

18/03/2017	-1	no change	no change	/	non-growing season	increase K and NG	0,091	0,226	-0,135
23/04/2017	-1	no change	no change	/	non-growing season	NG higher than K -> flow direction reversed	-0,010	0,140	-0,150
03/05/2017	-1	no change	normal increase	positive	start of growing season	NG higher than K -> flow direction reversed	-0,014	0,146	-0,160
07/05/2017	-1	no change	normal increase	positive	start of growing season	both curves flat	-0,009	0,116	-0,125
08/05/2017	-1	no change	normal increase	positive	start of growing season	both curves flat	-0,010	0,126	-0,136
12/05/2017	-1	normal increase	normal increase	/	growing season	both curves flat	-0,011	0,122	-0,133
13/05/2017	-1	normal increase	normal increase	/	growing season	both curves flat	-0,010	0,195	-0,205
17/05/2017	-1	normal increase	normal increase	/	growing season	both curves flat	-0,009	0,122	-0,131
18/05/2017	-1	normal increase	normal increase	/	growing season	NG higher than K -> flow direction reversed	-0,010	0,160	-0,170
02/07/2017	-1	normal increase	normal increase	/	growing season	NG higher than K -> flow direction reversed	-0,011	0,298	-0,309
07/07/2017	-1	normal increase	normal increase	/	growing season	just before return to original flow direction	-0,010	0,232	-0,242
16/07/2017	-1	normal increase	normal increase	/	growing season	NG higher than K -> flow direction reversed	0,002	0,186	-0,184
19/07/2017	-1	normal increase	normal increase	/	growing season	just before return to original flow direction	-0,011	0,151	-0,162
20/07/2017	-1	normal increase	small period of no change (minor influence)	negative	probably continuous growth	just before return to original flow direction	-0,012	0,175	-0,187
27/07/2017	1	small period of no change	normal increase	positive	probably continuous growth	K increase and NG decrease	0,238	0,072	0,166
30/08/2017	-1	normal increase	normal increase	/	growing season	just before return to original flow direction	-0,010	0,202	-0,212

02/09/2017	-1	normal increase	steep decrease	negative	mowing	/	0,065	0,221	-0,156
05/09/2017	-1	normal increase	steep decrease	negative	mowing	/	0,071	0,203	-0,132
09/09/2017	-1	steep decrease	steep decrease	/	mowing	strong increase K and NG	0,116	0,349	-0,233
22/11/2017	1	no change	no change	/	non-growing season	small peak K and NG	0,444	0,299	0,145
02/05/2018	1	no change	no change	/	growing season	/	0,576	0,431	0,145
10/05/2018	-1	no change	no change	/	growing season	/	0,043	0,183	-0,140
13/05/2018	-1	no change	no change	/	growing season	/	0,047	0,184	-0,137
29/05/2018	-1	no change	no change	/	growing season	/	0,111	0,239	-0,128
02/06/2018	1	no change	no change	/	growing season	increase K and NG	0,352	0,227	0,125
09/06/2018	1	no change	no change	/	growing season	K increase and NG decrease	0,420	0,170	0,250
23/06/2018	-1	no change	no change	/	growing season	strong decrease K	0,015	0,173	-0,158
08/09/2018	1	no change	no change	/	end of growing season	small peak K and NG	0,294	0,158	0,136
11/12/2018	1	no change	no change	/	non-growing season	K increase and NG decrease	0,451	0,293	0,158
14/06/2019	1	normal increase	steeper increase	positive	growing season	strong increase K and decrease NG	0,592	0,281	0,311
19/06/2019	-1	normal increase	normal increase	/	growing season	strong decrease K	0,085	0,302	-0,217
20/06/2019	-1	normal increase	normal increase	/	growing season	increase K	0,123	0,314	-0,191
06/09/2019	1	steep increase	no change	negative	end of growing season	/	0,354	0,078	0,276
07/09/2019	1	steep increase	no change	negative	end of growing season	/	0,342	0,113	0,229
18/09/2019	1	no change	no change	/	end of growing season	/	0,349	0,225	0,124
22/09/2019	1	no change	steep decrease	negative	mowing	/	0,291	0,171	0,120
29/09/2019	1	no change	steep decrease	negative	mowing	strong increase NG	0,512	0,287	0,225

Appendix 2: The anomalies of the gradient boosting model for the upper Amerongerwetering

date	residual factor	current vegetation state	next vegetation state	relative change in vegetation state	interpretation vegetation	particularities water level Amerongerwetering	observed gradient of the Amerongerwetering	modelled gradient of the Amerongerwetering	residual
12/03/2014	1	normal increase	normal increase	/	non-growing season	NG increase	0,362	0,237	0,125
16/06/2014	1	normal increase	no change	negative	growing season	/	0,302	0,103	0,199
06/07/2014	1	no change	no change	/	growing season	K increase	0,289	0,109	0,180
19/08/2014	1	data gap	data gap	/	growing season	K increase and NG decrease	0,474	0,309	0,165
24/09/2014	1	normal increase	no change	negative	end of growing season	/	0,343	0,197	0,146
29/09/2014	1	normal increase	no change	negative	end of growing season	/	0,397	0,272	0,125
11/10/2014	1	no change	normal decrease	negative	mowing	/	0,481	0,339	0,142
11/11/2014	-1	steadily increase	steadily increase	/	non-growing season	/	0,115	0,251	-0,136
12/11/2014	-1	steadily increase	steadily increase	/	non-growing season	/	0,112	0,239	-0,127
19/12/2014	-1	steadily increase	steadily increase	/	non-growing season	K decrease and NG increase	0,136	0,293	-0,157
24/12/2014	1	steadily increase	steadily increase	/	non-growing season	large peak K	0,392	0,270	0,122
16/03/2015	-1	steadily increase	steadily increase	/	non-growing season	increase NG	0,117	0,262	-0,145
22/03/2015	-1	steadily increase	steadily increase	/	non-growing season	/	0,146	0,269	-0,123
08/04/2015	-1	steadily increase	steadily increase	/	growing season	low peak K	0,137	0,269	-0,132
28/05/2015	-1	steep increase	normal increase	negative	growing season	decrease K	0,026	0,189	-0,163
19/07/2015	-1	steep increase	normal increase	negative	growing season	K and NG almost same value	0,053	0,204	-0,151
01/05/2016	1	possible data gap	steep increase	positive	growing season	/	0,374	0,227	0,147

02/05/2016	1	possible data gap	steep increase	positive	growing season	/	0,356	0,213	0,143
04/05/2016	1	possible data gap	steep increase	positive	growing season	/	0,347	0,219	0,128
01/06/2016	1	normal increase	no change	negative	growing season	K increase	0,516	0,343	0,173
14/06/2016	1	no change	normal decrease	negative	mowing	K increase	0,562	0,334	0,228
30/07/2016	1	no change	no change	/	growing after mowing	increase K	0,368	0,222	0,146
31/07/2016	1	no change	no change	/	growing season	increase K	0,442	0,311	0,131
10/08/2016	1	no change	steep decrease	negative	mowing	/	0,518	0,211	0,307
11/08/2016	1	no change	steep decrease	negative	mowing	/	0,555	0,289	0,266
15/08/2016	1	no change	steep decrease	negative	mowing	decrease K	0,367	0,203	0,164
26/08/2016	1	normal decrease	normal increase	positive	growing after mowing	NG decrease	0,288	0,117	0,171
02/09/2016	1	normal increase	normal increase	/	growing season	NG decrease	0,448	0,231	0,217
03/09/2016	1	normal increase	normal increase	/	growing season	both curves flat	0,448	0,230	0,218
04/09/2016	1	normal increase	normal increase	/	growing season	/	0,448	0,215	0,233
24/11/2016	-1	no change	no change	/	non-growing season	/	0,129	0,309	-0,180
26/02/2017	-1	no change	no change	/	non-growing season	/	0,184	0,342	-0,158
27/02/2017	-1	no change	no change	/	non-growing season	/	0,168	0,332	-0,164
01/03/2017	-1	no change	no change	/	non-growing season		0,244	0,390	-0,146
04/03/2017	-1	no change	no change	/	non-growing season	/	0,179	0,365	-0,186
06/03/2017	-1	no change	no change	/	non-growing season	/	0,210	0,376	-0,166
09/03/2017	-1	no change	no change	/	non-growing season	increase K and increase NG	0,279	0,451	-0,172

18/03/2017	-1	no change	no change	/	non-growing season	just before increase K	0,091	0,247	-0,156
23/04/2017	-1	no change	no change	/	non-growing season	NG higher than K -> flow direction reversed	-0,010	0,174	-0,184
12/05/2017	-1	normal increase	normal increase	/	growing season	both curves flat	-0,011	0,134	-0,145
13/05/2017	-1	normal increase	normal increase	/	growing season	NG higher than K -> flow direction reversed	-0,010	0,200	-0,210
17/05/2017	-1	normal increase	normal increase	/	growing season	NG higher than K -> flow direction reversed	-0,009	0,164	-0,173
18/05/2017	-1	normal increase	normal increase	/	growing season	NG higher than K -> flow direction reversed	-0,010	0,146	-0,156
02/07/2017	-1	normal increase	normal increase	/	growing season	NG higher than K -> flow direction reversed	-0,011	0,278	-0,289
07/07/2017	-1	normal increase	normal increase	/	growing season	just before return to original flow direction	-0,010	0,233	-0,243
16/07/2017	-1	normal increase	normal increase	/	growing season	/	0,002	0,144	-0,142
19/07/2017	-1	normal increase	normal increase	/	growing season	just before return to original flow direction	-0,011	0,145	-0,156
20/07/2017	-1	normal increase	normal increase	/	growing season	just before return to original flow direction	-0,012	0,141	-0,153
27/07/2017	1	normal increase	normal increase	/	growing season	K increase and NG decrease	0,238	0,045	0,193
30/08/2017	-1	normal increase	steep decrease	negative	mowing	just before return to original flow direction	-0,010	0,147	-0,157
02/09/2017	-1	normal increase	steep decrease	negative	mowing	increase K and increase NG	0,065	0,215	-0,150
05/09/2017	-1	normal increase	steep decrease	negative	mowing	decrease K and decrease NG	0,071	0,196	-0,125
09/09/2017	-1	steep decrease	steep decrease	/	mowing	strong increase K and NG	0,116	0,437	-0,321
08/10/2017	-1	no change	no change	/	non-growing season	decrease K and increase NG	0,253	0,423	-0,170
22/11/2017	1	no change	no change	/	non-growing season	small peak K and NG	0,444	0,316	0,128
02/05/2018	1	no change	no change	/	growing season	decrease K and decrease NG	0,576	0,371	0,205
29/05/2018	-1	no change	no change	/	growing season	/	0,111	0,283	-0,172

02/06/2018	1	no change	no change	/	growing season	increase K and increase NG	0,352	0,229	0,123
07/06/2018	1	no change	no change	/	growing season	/	0,206	0,079	0,127
09/06/2018	1	no change	no change	/	growing season	K increase and NG decrease	0,420	0,206	0,214
13/06/2018	1	no change	no change	/	growing season	/	0,234	0,093	0,141
08/09/2018	1	no change	no change	/	end of growing season	small peak K and NG	0,294	0,148	0,146
11/12/2018	1	no change	no change	/	non-growing season	decrease K and decrease NG	0,451	0,321	0,130
14/06/2019	1	normal increase	steeper increase	positive	growing season	strong increase K and decrease NG	0,592	0,302	0,290
19/06/2019	-1	normal increase	normal increase	/	growing season	strong decrease K	0,085	0,258	-0,173
20/06/2019	-1	normal increase	normal increase	/	growing season	increase K	0,123	0,291	-0,168
22/06/2019	-1	steep increase	steep increase	/	mowing	decrease K	0,088	0,223	-0,135
06/09/2019	1	steep increase	no change	negative	end of growing season	/	0,354	0,078	0,276
07/09/2019	1	steep increase	no change	negative	end of growing season	/	0,342	0,156	0,186
29/09/2019	1	steep decrease	steep decrease	/	mowing	strong increase NG	0,512	0,269	0,243

Appendix 3: The anomalies of the random forest model for the lower Amerongerwetering

date	residual factor	current vegetation state	next vegetation state	relative change in vegetation state	interpretation vegetation	particularities water level Amerongerwetering	observed gradient of the Amerongerwetering	modelled gradient of the Amerongerwetering	residual
11/04/2014	-1	steadily increase	steadily increase	/	non-growing season	/	0,085	0,214	-0,129
29/04/2014	-1	steadily increase	steadily increase	/	non-growing season	/	0,088	0,286	-0,198
30/04/2014	-1	steadily increase	steadily increase	/	non-growing season	/	0,090	0,239	-0,149
10/07/2014	1	no change	steep increase	positive	growing season	NG increase	0,312	0,159	0,153
15/08/2014	-1	data gap	data gap	/	growing season	/	0,089	0,236	-0,147
05/09/2014	-1	data gap	data gap	/	growing season	/	0,090	0,272	-0,182
16/08/2015	-1	data gap	data gap	/	growing season	NG increase	0,230	0,356	-0,126
19/08/2015	1	normal increase	normal increase	/	growing season	NG increase	0,541	0,284	0,257
04/09/2015	1	normal increase	no change	negative	growing season	NG decrease and AW increase	0,623	0,444	0,179
05/09/2015	-1	normal increase	no change	negative	growing season	NG decrease and AW increase	0,494	0,652	-0,158
06/09/2015	1	normal increase	no change	negative	growing season	NG increase and AW decrease	0,910	0,712	0,198
07/09/2015	1	normal increase	no change	negative	growing season	NG increase and AW decrease	0,903	0,693	0,210
18/09/2015	1	no change	steep increase	positive	growing season	/	0,598	0,410	0,188
21/09/2015	1	no change	steep increase	positive	growing season	/	0,712	0,393	0,319
22/09/2015	1	no change	steep increase	positive	growing season	NG decrease	0,611	0,457	0,154
17/10/2015	1	steep increase	normal decrease	negative	mowing	NG increase	0,610	0,432	0,178
22/02/2016	1	possible data gap	possible data gap	/	non-growing season	NG increase and AW decrease	0,677	0,404	0,273

26/02/2016	-1	possible data gap	possible data gap	/	non-growing season	NG decrease	0,239	0,371	-0,132
02/04/2016	-1	possible data gap	possible data gap	/	non-growing season	/	0,085	0,204	-0,119
08/04/2016	-1	possible data gap	possible data gap	/	non-growing season	/	0,091	0,231	-0,140
25/04/2016	-1	possible data gap	possible data gap	/	non-growing season	/	0,084	0,230	-0,146
01/05/2016	-1	possible data gap	steep increase	positive	growing season	/	0,123	0,241	-0,118
02/05/2016	-1	possible data gap	steep increase	positive	growing season	/	0,099	0,236	-0,137
03/05/2016	-1	possible data gap	steep increase	positive	growing season	NG and AW increasing	0,094	0,241	-0,147
21/05/2016	1	steep increase	normal increase	negative	growing season	/	0,512	0,314	0,198
01/06/2016	1	normal increase	no change	negative	growing season	NG increase	0,471	0,291	0,180
15/06/2016	1	no change	normal decrease	negative	mowing	NG increase and AW decrease	0,758	0,574	0,184
18/06/2016	1	no change	normal decrease	negative	mowing	NG increase and AW decrease	0,827	0,539	0,288
25/06/2016	-1	normal decrease	normal decrease	/	mowing	NG decrease and AW increase	0,354	0,512	-0,158
24/02/2017	1	no change	no change	/	non-growing season	NG increase	0,586	0,429	0,157
17/03/2017	1	no change	no change	/	non-growing season	/	0,429	0,139	0,290
18/03/2017	1	no change	no change	/	non-growing season	/	0,424	0,184	0,240
11/09/2017	1	steep decrease	steep decrease	/	mowing	NG decrease and AW increase	0,831	0,581	0,250
17/09/2017	1	steep decrease	less steep decrease	negative	mowing	NG at high peak and AW at low peak	0,843	0,590	0,253
18/09/2017	1	steep decrease	less steep decrease	negative	mowing	NG decrease and AW increase	0,812	0,610	0,202
06/04/2018	1	no change	no change	/	non-growing season	small increase peak of both NG and AW	0,380	0,184	0,196
01/05/2018	1	no change	no change	/	growing season	NG at high peak and AW at low peak	0,761	0,422	0,339

16/05/2018	1	no change	no change	/	growing season	small increase peak NG and small decrease peak of AW	0,387	0,230	0,157
03/06/2018	1	no change	no change	/	growing season	NG increase and AW decrease	0,492	0,288	0,204
08/08/2018	1	steep increase	no change	negative	end of growing season	AW increase after low peak	0,605	0,403	0,202
07/03/2019	-1	no change	no change	/	non-growing season	NG increase after low peak	0,092	0,253	-0,161
25/03/2019	-1	no change	no change	/	non-growing season	NG decrease	0,143	0,271	-0,128
04/04/2019	-1	no change	no change	/	non-growing season	NG and AW increasing	0,093	0,269	-0,176
24/04/2019	-1	no change	no change	/	non-growing season	/	0,043	0,192	-0,149
08/05/2019	-1	no change	no change	/	growing season	small decrease peak of both NG and AW	0,063	0,238	-0,175
09/05/2019	-1	no change	no change	/	growing season	/	0,060	0,224	-0,164
10/05/2019	-1	no change	no change	/	growing season	/	0,069	0,201	-0,132
11/05/2019	-1	no change	no change	/	growing season	/	0,066	0,191	-0,125
14/06/2019	1	normal increase	steep increase	positive	growing season	NG increase peak and AW decrease peak	0,513	0,280	0,233
26/09/2019	-1	no change	steep decrease	negative	mowing	NG increase and AW decrease	0,171	0,307	-0,136
27/09/2019	-1	no change	steep decrease	negative	mowing	NG increase and AW decrease	0,199	0,320	-0,121
02/10/2019	1	steep decrease	no change	positive	mowing	NG increase peak and AW decrease peak	0,804	0,525	0,279
03/10/2019	1	steep decrease	no change	positive	mowing	NG decreasing and AW increasing after peaks	0,784	0,584	0,200
04/10/2019	1	steep decrease	no change	positive	mowing	NG decreasing and AW increasing after peaks	0,441	0,222	0,219
13/11/2019	-1	no change	no change	/	non-growing season	/	0,226	0,375	-0,149
27/11/2019	-1	no change	no change	/	non-growing season	just before NG high peak	0,138	0,264	-0,126
02/12/2019	-1	no change	no change	/	non-growing season	NG decrease	0,165	0,307	-0,142
07/12/2019	-1	no change	no change	/	non-growing season	NG increase	0,235	0,370	-0,135

Appendix 4: The anomalies of the gradient boosting model for the lower Amerongerwetering

date	residual factor	current vegetation state	next vegetation state	relative change in vegetation state	interpretation vegetation	particularities water level Amerongerwetering	observed gradient of the Amerongerwetering	modelled gradient of the Amerongerwetering	residual
11/04/2014	-1	steadily increase	steadily increase	/	non-growing season	/	0,085	0,267	-0,182
29/04/2014	-1	steadily increase	steadily increase	/	non-growing season	/	0,088	0,329	-0,241
30/04/2014	-1	steadily increase	steadily increase	/	non-growing season	/	0,090	0,271	-0,181
01/05/2014	-1	steadily increase	steadily increase	/	growing season	/	0,095	0,250	-0,155
20/10/2014	-1	normal decrease	steadily increase	positive	non-growing season	just before high peak of NG	0,115	0,279	-0,164
30/10/2014	-1	normal decrease	steadily increase	positive	non-growing season	just after high peak of NG	0,099	0,268	-0,169
13/12/2014	1	steadily increase	steadily increase	/	non-growing season	high peak of NG	0,477	0,268	0,209
19/08/2015	1	normal increase	normal increase	/	growing season	NG increase	0,541	0,220	0,321
04/09/2015	1	normal increase	no change	negative	growing season	NG decrease and AW increase	0,623	0,385	0,238
05/09/2015	-1	normal increase	no change	negative	growing season	NG decrease and AW increase	0,494	0,699	-0,205
06/09/2015	1	normal increase	no change	negative	growing season	NG increase and AW decrease	0,910	0,741	0,169
21/09/2015	1	steep increase	steep increase	/	growing season	/	0,712	0,330	0,382
22/09/2015	1	steep increase	steep increase	/	growing season	NG decrease	0,611	0,360	0,251
23/09/2015	1	steep increase	steep increase	/	growing season	NG increase	0,635	0,292	0,343
30/09/2015	1	steep increase	steep increase	/	growing season	NG decrease and AW increase	0,392	0,224	0,168
01/10/2015	1	steep increase	steep increase	/	growing season	/	0,370	0,179	0,191

17/10/2015	1	steep increase	normal decrease	negative	mowing	NG increase	0,610	0,459	0,151
16/11/2015	1	possible data gap	possible data gap	/	non-growing season	high peak of NG	0,535	0,340	0,195
13/12/2015	1	possible data gap	possible data gap	/	non-growing season	high peak of NG	0,440	0,256	0,184
19/01/2016	-1	possible data gap	possible data gap	/	non-growing season	NG decrease	0,158	0,308	-0,150
08/03/2016	-1	possible data gap	possible data gap	/	non-growing season	NG decrease	0,254	0,403	-0,149
04/04/2016	-1	possible data gap	possible data gap	/	non-growing season	/	0,076	0,251	-0,175
08/04/2016	-1	possible data gap	possible data gap	/	non-growing season	/	0,091	0,302	-0,211
09/04/2016	-1	possible data gap	possible data gap	/	non-growing season	/	0,085	0,256	-0,171
10/04/2016	-1	possible data gap	possible data gap	/	non-growing season	/	0,079	0,253	-0,174
12/04/2016	-1	possible data gap	possible data gap	/	non-growing season	/	0,067	0,256	-0,189
13/04/2016	-1	possible data gap	possible data gap	/	non-growing season	NG and AW decrease	0,075	0,232	-0,157
14/04/2016	-1	possible data gap	possible data gap	/	non-growing season	/	0,082	0,265	-0,183
18/04/2016	-1	possible data gap	possible data gap	/	non-growing season	/	0,092	0,292	-0,200
20/04/2016	-1	possible data gap	possible data gap	/	non-growing season	/	0,089	0,286	-0,197
21/04/2016	-1	possible data gap	possible data gap	/	non-growing season	/	0,081	0,283	-0,202
23/04/2016	-1	possible data gap	possible data gap	/	non-growing season	/	0,081	0,234	-0,153
24/04/2016	-1	possible data gap	possible data gap	/	non-growing season	/	0,081	0,270	-0,189
25/04/2016	-1	possible data gap	possible data gap	/	non-growing season	/	0,084	0,254	-0,170
26/04/2016	-1	possible data gap	possible data gap	/	non-growing season	NG increasing	0,092	0,349	-0,257
01/05/2016	-1	possible data gap	possible data gap	/	growing season	NG decreasing	0,123	0,347	-0,224

02/05/2016	-1	possible data gap	possible data gap	/	growing season	NG decreasing	0,099	0,340	-0,241
03/05/2016	-1	normal decrease	steep increase	positive	growing season	/	0,094	0,317	-0,223
04/05/2016	-1	normal decrease	steep increase	positive	growing season	NG and AW increasing	0,097	0,260	-0,163
05/05/2016	-1	normal decrease	steep increase	positive	growing season	NG and AW increasing	0,080	0,244	-0,164
09/05/2016	-1	steep increase	steep increase	/	growing season	datagap NG	0,071	0,246	-0,175
16/05/2016	1	steep increase	steep increase	/	growing season	just after data gap	0,472	0,330	0,142
18/05/2016	1	steep increase	normal increase	negative	growing season	/	0,476	0,293	0,183
01/06/2016	1	normal increase	no change	negative	growing season	increase NG and decrease AW	0,471	0,336	0,135
19/06/2016	1	steep increase	normal increase	negative	growing season	just after high peak of NG	0,797	0,508	0,289
06/02/2017	1	no change	no change	/	non-growing season	/	0,404	0,270	0,134
17/03/2017	1	no change	no change	/	non-growing season	/	0,429	0,188	0,241
21/03/2017	1	no change	no change	/	non-growing season	/	0,413	0,273	0,140
27/07/2017	1	normal increase	normal increase	/	growing season	NG increase and AW decrease	0,435	0,301	0,134
06/10/2017	1	steep decrease	no change	positive	non-growing season	NG high peak and small AW low peak	0,625	0,409	0,216
09/06/2018	1	no change	no change	/	growing season	small NG high peak and small AW low peak	0,428	0,233	0,195
10/06/2018	1	no change	no change	/	growing season	small NG high peak and small AW low peak	0,457	0,220	0,237
08/02/2019	1	no change	no change	/	non-growing season	high peak of NG	0,506	0,350	0,156
28/09/2019	1	steep decrease	steep decrease	/	mowing	high peak of NG and low peak of AW	0,495	0,185	0,310
02/10/2019	1	steep decrease	no change	positive	mowing	NG increase peak and AW decrease peak	0,804	0,545	0,259
03/10/2019	1	steep decrease	no change	positive	mowing	NG decreasing and AW increasing after peaks	0,784	0,567	0,217

04/10/2019	1	steep decrease	no change	positive	mowing	NG decrease and AW increase	0,441	0,304	0,137
03/11/2019	1	data gap	data gap	/	/	high peak of NG and low peak of AW	0,475	0,302	0,173

Appendix 5: The anomalies of the random forest model for the Lange Weide drainage area

date	residual factor	current vegetation state	next vegetation state	relative change in vegetation state	interpretation vegetation	particularities water level Lange Weide	observed gradient of Lange Weide	modelled gradient of Lange Weide	residual
12/05/2014	1	no data	no data	/	/	/	0,078	-0,024	0,102
10/07/2014	1	no data	no data	/	/	/	0,125	0,043	0,082
28/07/2014	-1	no data	no data	/	/	just before positive peak IN and PS	-0,036	0,126	-0,162
31/07/2014	-1	no data	no data	/	/	just after positive peak IN and PS	-0,008	0,034	-0,042
02/08/2014	-1	no data	no data	/	/	IN and PS almost same value	-0,008	0,037	-0,045
17/08/2014	-1	no data	no data	/	/	IN and PS almost same value	0,017	0,082	-0,065
27/09/2014	-1	no change	no change	/	end of growing season	just before positive peak IN and PS, PS positive peak higher	-0,049	0,004	-0,053
12/10/2014	-1	no change	no change	/	non-growing season	IN and PS almost same value	-0,007	0,044	-0,051
17/10/2014	-1	no change	slow increase	positive	non-growing season	IN and PS almost same value	-0,178	0,048	-0,226
18/10/2014	-1	no change	slow increase	positive	non-growing season	just before large negative peak IN	-0,651	-0,028	-0,623
19/10/2014	-1	no change	slow increase	positive	non-growing season	large negative peak IN	-0,612	-0,044	-0,568
25/10/2014	-1	slow increase	slow increase	/	non-growing season	large negative peak IN	-0,622	-0,032	-0,590
17/11/2014	-1	slow increase	slow increase	/	non-growing season	/	0,011	0,056	-0,045
21/01/2015	-1	slow increase	slow increase	/	non-growing season	IN and PS almost same value	-0,038	0,000	-0,038
10/03/2015	-1	slow increase	slow increase	/	non-growing season	PS larger than IN	-0,035	0,006	-0,041
29/03/2015	-1	slow increase	slow increase	/	non-growing season	/	0,032	0,069	-0,037
06/04/2015	-1	slow increase	slow increase	/	non-growing season	IN and PS almost same value	-0,021	0,016	-0,037
06/09/2015	1	steep decrease	less steep decrease	positive	mowing	large positive peak IN, data gap PS	0,195	0,102	0,093

17/11/2015	1	no change	no change	/	non-growing season	positive peak IN, negative peak PS	0,195	0,099	0,096
29/11/2015	1	no change	no change	/	non-growing season	just before positive peak IN and small negative peak PS	0,210	0,090	0,120
20/02/2016	1	no change	no change	/	non-growing season	just before positive peak IN and small negative peak PS	0,154	0,064	0,090
02/06/2016	1	steep increase	steep increase	/	growing season	/	0,103	0,014	0,089
15/06/2016	1	steep increase	steep increase	/	growing season	just before large positive peak IN and large negative peak PS	0,244	0,031	0,213
16/06/2016	1	steep increase	steep increase	/	growing season	large positive peak IN and large negative peak PS	0,356	0,111	0,245
17/06/2016	1	steep increase	steep increase	/	growing season	large positive peak IN and large negative peak PS	0,313	0,100	0,213
18/06/2016	1	steep increase	steep increase	/	growing season	large positive peak IN and large negative peak PS	0,226	0,073	0,153
27/06/2016	-1	steep increase	steep increase	/	growing season	just after large positive peak IN and negative peak PS	0,068	0,111	-0,043
01/07/2016	-1	steep increase	steep increase	/	growing season	/	0,053	0,094	-0,041
08/07/2016	-1	steep increase	less steep increase	negative	growing season	/	0,043	0,082	-0,039
09/08/2016	1	no change	no change	/	end of growing season	large peak IN and negative peak PS	0,207	0,107	0,100
10/08/2016	1	no change	no change	/	end of growing season	large peak IN and negative peak PS	0,229	0,090	0,139
11/08/2016	1	no change	no change	/	end of growing season	large peak IN and negative peak PS	0,226	0,087	0,139
14/08/2016	1	no change	no change	/	end of growing season	large peak IN and negative peak PS	0,225	0,143	0,082
20/08/2016	1	no change	no change	/	end of growing season	large peak IN and negative peak PS	0,214	0,128	0,086
17/11/2016	1	normal decrease	normal decrease	/	non-growing season	just before large positive peak IN and large negative peak PS	0,221	0,088	0,133
22/06/2017	1	normal increase	normal increase	/	growing season	just before positive peak IN and negative peak PS	0,171	0,057	0,114
14/07/2017	-1	normal increase	normal increase	/	growing season	just after positive peak IN and negative peak PS	0,032	0,071	-0,039
20/07/2017	-1	normal increase	normal increase	/	growing season	/	0,050	0,092	-0,042

29/07/2017	1	normal increase	normal increase	/	growing season	just before positive peak IN and negative peak PS	0,176	0,094	0,082
30/07/2017	1	normal increase	normal increase	/	growing season	just before positive peak IN and negative peak PS	0,141	0,037	0,104
15/08/2017	1	no change	normal increase	positive	growing season	positive peak IN and negative peak PS	0,217	0,082	0,135
16/08/2017	1	no change	normal increase	positive	growing season	positive peak IN and negative peak PS	0,164	0,082	0,082
15/09/2017	1	no change	no change	/	end of growing season	just before positive peak IN and negative peak PS	0,217	0,107	0,110
16/09/2017	1	no change	no change	/	end of growing season	just before positive peak IN and negative peak PS	0,210	0,109	0,101
21/09/2017	-1	no change	no change	/	end of growing season	/	0,030	0,079	-0,049
06/10/2017	1	no change	no change	/	non-growing season	just before positive peak IN and negative peak PS	0,123	0,005	0,118
08/10/2017	1	no change	no change	/	non-growing season	just before positive peak IN and negative peak PS	0,226	0,012	0,214
08/12/2017	1	normal decrease	normal decrease	/	non-growing season	just before large positive peak IN and negative peak PS	0,241	0,108	0,133
12/12/2017	1	normal decrease	normal decrease	/	non-growing season	just before large positive peak IN and large positive peak PS	0,126	0,050	0,076
24/01/2018	-1	normal decrease	normal decrease	/	non-growing season	/	0,014	0,051	-0,037
16/05/2018	-1	no change	no change	/	start of growing season	IN and PS almost same value	-0,004	0,042	-0,046
31/05/2018	1	no change	steep increase	positive	growing season	just before positive peak IN and negative peak PS	0,174	0,085	0,089
02/06/2018	1	no change	steep increase	positive	growing season	positive peak IN and negative peak PS	0,144	0,066	0,078
09/09/2018	-1	no change	no change	/	end of growing season	just after large positive peak IN and positive peak PS	0,020	0,072	-0,052
15/09/2018	-1	no change	steep decrease	negative	mowing	/	0,011	0,048	-0,037
08/06/2019	-1	no change	steep increase	positive	growing season	large positive peak IN and large negative peak PS	0,014	0,059	-0,045
12/07/2019	-1	no change	no change	/	end of growing season	/	0,043	0,096	-0,053
26/09/2019	-1	no change	normal decrease	negative	mowing	IN and PS almost same value	0,043	0,090	-0,047

29/09/2019	-1	no change	normal decrease	negative	mowing	decrease IN and PS	0,028	0,123	-0,095
30/09/2019	-1	no change	normal decrease	negative	mowing	/	0,042	0,087	-0,045

Appendix 6: The anomalies of the gradient boosting model for the Lange Weide drainage area

date	residual factor	current vegetation state	next vegetation state	relative change in vegetation state	interpretation vegetation	particularities water level Lange Weide	observed gradient of Lange Weide	modelled gradient of Lange Weide	residual
12/05/2014	1	no data	no data	/	/	/	0,078	-0,087	0,165
16/05/2014	1	no data	no data	/	/	/	0,051	-0,030	0,081
26/05/2014	1	no data	no data	/	/	/	0,082	-0,001	0,083
28/07/2014	-1	no data	no data	/	/	just before positive peak IN and PS	-0,036	0,111	-0,147
17/08/2014	-1	no data	no data	/	/	IN and PS almost same value	0,017	0,086	-0,069
27/09/2014	-1	no change	no change	/	end of growing season	just before positive peak IN and PS, PS positive peak higher	-0,049	0,016	-0,065
01/10/2014	-1	no change	no change	/	non-growing season	/	-0,022	0,030	-0,052
12/10/2014	-1	no change	no change	/	non-growing season	IN and PS almost same value	-0,007	0,062	-0,069
17/10/2014	-1	no change	slow increase	positive	non-growing season	IN and PS almost same value	-0,178	0,029	-0,207
18/10/2014	-1	no change	slow increase	positive	non-growing season	just before large negative peak IN	-0,651	-0,088	-0,563
19/10/2014	-1	no change	slow increase	positive	non-growing season	large negative peak IN	-0,612	-0,149	-0,463
25/10/2014	-1	slow increase	slow increase	/	non-growing season	large negative peak IN	-0,622	-0,123	-0,499
17/11/2014	-1	slow increase	slow increase	/	non-growing season	/	0,011	0,069	-0,058
21/01/2015	-1	slow increase	slow increase	/	non-growing season	IN and PS almost same value	-0,038	0,007	-0,045
29/01/2015	-1	slow increase	slow increase	/	non-growing season	IN and PS almost same value	0,040	0,092	-0,052
06/09/2015	1	steep increase	steep increase	/	mowing	large positive peak IN, data gap PS	0,195	0,071	0,124
10/09/2015	1	steep increase	steep increase	/	mowing	just before large negative peak IN	0,047	-0,104	0,151
20/09/2015	1	steep increase	normal increase	negative	mowing	just after positive peak IN	0,053	-0,047	0,100

14/11/2015	1	no change	no change	/	non-growing season	/	0,130	0,041	0,089
29/11/2015	1	no change	no change	/	non-growing season	just before positive peak IN and small negative peak PS	0,210	0,120	0,090
31/01/2016	-1	no change	no change	/	non-growing season	negative peak PS	0,064	0,113	-0,049
20/02/2016	1	no change	no change	/	non-growing season	just before positive peak IN and small negative peak PS	0,154	0,073	0,081
15/06/2016	1	steep increase	steep increase	/	growing season	just before large positive peak IN and large negative peak PS	0,244	0,024	0,220
16/06/2016	1	steep increase	steep increase	/	growing season	large positive peak IN and large negative peak PS	0,356	0,165	0,191
17/06/2016	1	steep increase	steep increase	/	growing season	large positive peak IN and large negative peak PS	0,313	0,123	0,190
18/06/2016	1	steep increase	steep increase	/	growing season	large positive peak IN and large negative peak PS	0,226	0,087	0,139
27/06/2016	-1	steep increase	steep increase	/	growing season	just after large positive peak IN and negative peak PS	0,068	0,134	-0,066
01/07/2016	-1	steep increase	steep increase	/	growing season	/	0,053	0,104	-0,051
04/07/2016	-1	steep increase	less steep increase	negative	growing season	/	0,056	0,115	-0,059
09/07/2016	-1	steep increase	less steep increase	negative	growing season	/	0,031	0,088	-0,057
05/08/2016	-1	no change	no change	/	end of growing season	large positive peak IN	0,133	0,188	-0,055
11/08/2016	1	no change	no change	/	end of growing season	large peak IN and negative peak PS	0,226	0,136	0,090
13/10/2016	1	no change	no change	/	non-growing season	/	0,049	-0,098	0,147
15/10/2016	1	no change	no change	/	non-growing season	/	0,063	-0,080	0,143
17/11/2016	1	normal decrease	normal decrease	/	non-growing season	just before large positive peak IN and large negative peak PS	0,221	0,066	0,155
23/11/2016	-1	possible data gap	possible data gap	/	non-growing season	/	0,009	0,057	-0,048
12/01/2017	1	possible data gap	possible data gap	/	non-growing season	/	0,122	0,042	0,080
22/06/2017	1	steep increase	steep increase	/	growing season	just before positive peak IN and negative peak PS	0,171	0,064	0,107

14/07/2017	-1	normal increase	normal increase	/	growing season	just after positive peak IN and negative peak PS	0,032	0,089	-0,057
29/07/2017	1	normal increase	normal increase	/	growing season	just before positive peak IN and negative peak PS	0,176	0,067	0,109
30/07/2017	1	normal increase	normal increase	/	growing season	just before positive peak IN and negative peak PS	0,141	0,020	0,121
15/08/2017	1	no change	normal increase	positive	growing season	positive peak IN and negative peak PS	0,217	0,084	0,133
16/08/2017	1	no change	normal increase	positive	growing season	positive peak IN and negative peak PS	0,164	0,078	0,086
14/09/2017	1	steep decrease	no change	positive	mowing	just before large positive peak IN	0,186	0,104	0,082
18/09/2017	-1	steep decrease	no change	positive	mowing	just after large positive peak IN and negative peak PS	0,070	0,128	-0,058
21/09/2017	-1	no change	no change	/	end of growing season	/	0,030	0,114	-0,084
06/10/2017	1	no change	no change	/	non-growing season	just before positive peak IN and negative peak PS	0,123	0,027	0,096
08/10/2017	1	no change	no change	/	non-growing season	just before positive peak IN and negative peak PS	0,226	0,058	0,168
08/12/2017	1	normal decrease	normal decrease	/	non-growing season	just before large positive peak IN and negative peak PS	0,241	0,052	0,189
19/12/2017	-1	possible data gap	possible data gap	/	non-growing season	/	0,006	0,060	-0,054
15/01/2018	1	possible data gap	possible data gap	/	non-growing season	/	0,104	0,003	0,101
31/05/2018	1	no change	steep increase	positive	growing season	just before positive peak IN and negative peak PS	0,174	0,086	0,088
02/06/2018	1	no change	steep increase	positive	growing season	positive peak IN and negative peak PS	0,144	0,023	0,121
09/09/2018	-1	no change	no change	/	end of growing season	just after large positive peak IN and positive peak PS	0,020	0,075	-0,055
21/09/2018	-1	steep decrease	steep increase	/	mowing	/	0,048	0,097	-0,049
25/10/2018	-1	normal decrease	normal increase	positive	non-growing season	/	0,008	0,063	-0,055
12/07/2019	-1	no change	no change	/	end of growing season	/	0,043	0,113	-0,070
13/08/2019	-1	no change	no change	/	end of growing season	just before positive peak IN and negative peak PS	0,056	0,113	-0,057

26/09/2019	-1	no change	normal decrease	negative	mowing	IN and PS almost same value	0,043	0,103	-0,060
29/09/2019	-1	no change	normal decrease	negative	mowing	decrease IN and PS	0,028	0,148	-0,120