University of Groningen

Master's Design Project IEM (WMIE15002)

Decision Support Tool For Safe Downscaling of a Water Treatment Plant

Student:
Alexandru Olar (S3419169)

Supervisors:
Mehran Mohebbi (First)
Bayu Jayawardhanna (Second)
Robert Schots (Company)

March 19, 2021

Abstract

Water Laboratorium Noord (WLN) is a subsidiary of the Waterbedrijf Groningen (Water Company Groningen, WCG) tasked with performing water analysis, research on water related topics and act as a consultant for water technologies. WCG own the drinking water treatment plant at De Punt which intakes water from the Drentsche Aa river and produces drinking water for the city of Groningen and the surrounding area. During periods of drought, usually during the summer months, the supply of water available in the river drops to levels close or lower than the plants capacity. Failure to react to these drops, and downscale the capacity of the plant to match the supply of the river could lead to massive damage to plant components and downtime. This project aims to develop a Machine Learning (ML) model to predict the supply the water available in the river using meteorological data. In order to achieve this, data from multiple weather stations together with data from multiple SCADA systems managed by the WCG and Water Board Hunze en Aa's (WB) will be included in the analysis. Multiple ML models, including Linear Regression (LR), Decision Trees (DT) and various ensemble type methods, such as Random Forest (RF) will be used to predict the streamflow of the river with multiple lead times. Two models were developed to predict the streamflow in 6 and 12 days. New features were generated after which Forward and Exhaustive Feature Selection was used to obtain the optimal subset of features. The best performing model for both time periods was the RF. Shapley values for the best subset of features for both models were computed and presented. A software solution was developed that will update data to the current day, train both models using the most current dataset and predict the flow values for the next twelve days.

Contents

Abreviatons

WTP - Water Treatment Plant
WLN - Water Laboratorium Noord
WCG - Water Company Groningen
WB - Water Board
ML - Machine Learning
DS - Data Science
LR - Linear Regression
DT - Decision Tree
RF - Random Forest
GB - Gradient Boost
ET - Extra randomized Tees
AB - Ada Boost
XGB - EXtreme Gradient Boost
SW - Surface Water
GW - Ground Water
RB - Retention Basin
CS - Coagulation / Sedimentation
RSF - Rapid Sand FIltration
SSF - Slow Sand Filtration
ACF - Active Carbon FIltration
UVD - UV Disinfection
CA - Cascade Aeration
SI - Smart Industry
ENTEG - ENgineering and TEchnology institute Groningen
MI - Model Interpretation

1 Introduction

Waterbedrijf Groningen (Water Company Groningen, WCG) is responsible for supplying water to the city and province of Groningen as well as the former municipality of Eelde with clean drinking water. Established in 1989 through the merger of the Provincial Water Company and the Municipal Water Company of Gronigen, it produces around 44 million m^3 of drinking water out of the national 1150 million m^3, with a turnover rate of 45 million euros. Waterlaboratorium Noord (Water Laboratory North, WLN), a subsidiary of the WCG and the Waterleiding-Maatschappij Drenthe (Water Supply Company Drenthe, WMD) as well as being the executant of this project, is a consultancy company focusing of drinking water activities. It is responsible with performing chemical water sample analysis at the company's on site laboratory, provide technological advice for water related activities and perform research in the field of water treatment.

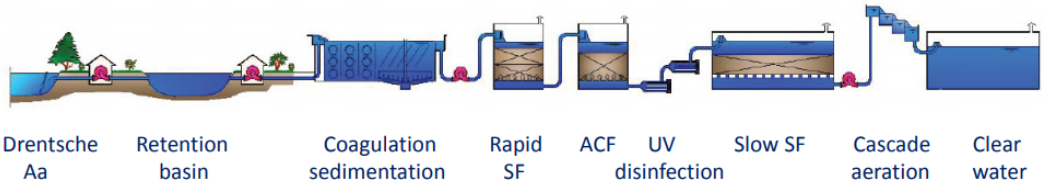

Figure 1: De Punt water treatment plant process diagram containing all the steps in the water treatment process

The water treatment plant (WTP) at De Punt (Figure 1) is designed to function at a maximum capacity of about 840 m^3/h in the intake. However during the summer months, periods of drought can lead to the supply of water in the river dropping to levels close to or lower than the maximum capacity of the plant. Lack of water in the inlet leads to the removal of the medium by which parts of the treatment process are performed, leading to process inefficiency. To combat this, in case of a supply shortage, water from the local ground water (GW) treatment plant is redirected to the outlet of the coagulation/sedimentation process while the capacity of the surface treatment plant is downscaled to match the supply from the river. Some of the treatment steps make us of various bacteria to clean the water, bacteria which are highly dependent on temperature and oxygen. The process of downscaling the plant is a time consuming and delicate issue, as during the summer water from the ground water is much colder than surface water, and a fast substitution to ground water could lead to significant water temperature and oxygen imbalances that could kill or inactivate most bacteria. As such any downscaling of the plant needs to be planned ahead and must be performed under a strict strategy.

1.1 Background

Water is one of the most essential resources to mankind. At the moment we harvest only about 0.08% of the entire planets fresh water capacity while the demand for fresh water for drinking, manufacturing, sanitation, agriculture and other human water dependent activities is ever increasing. Drinking water is water that is safe for human consumption or food preparation (World Health Organization 2017) and it must be treated before use. Sources of drinking water include surface waters, such as rivers, creeks or lakes,

and ground water, water present beneath the earth in soil pore spaces and fractures of rock formations. Surface water is responsible for about 68% of water that is provided to communities (Agency n.d.) and it is measured by the amount of runoff it generates, including sources such as rainfall and snow melting, and drains such as land, vegetation and water surface evapo-transpiration mechanisms.

Water Resource Management (WRM) is the activity of planning, developing, distributing and managing the optimum use of water resources, such as surface water. Successfully managing these water resources requires extensive knowledge on the available sources, the processes required to treat this water and the means by which it is delivered to the end users. One of the challenges water managers face is water scarcity and shortages. Surface water is naturally replenished by rainfall and lost through evapo-transpiration. However during periods of drought, there is a consistent reduction in the water availability expressed through rainfall, and an excess of heat resulting in more evaporation (Yevjevich, Cunha, and Vlachos 2012).

Droughts can be distinguished according to the hydrologic cycle which it affects (G. Rossi, Castiglione, and B. Bonaccorso 2007). A meteorological drought refers a condition of reduction in the amount of precipitation compared to normal values. This can lead to a soil moisture deficit which in turn can affect surface water and ground water bodies resulting in what is called a hydrological drought. Since these bodies of water represent the main sources of water for various water supply systems, a deficit in any of these two will lead to an operational drought which can have significant economic, environmental and social impacts. The economic effect of such a drought comes from the income reduction of water companies due to reduced water availability, environmental concerns refer to a lack of feed and drinking water and social impacts are caused by inconveniences due to water systems rationing.

Because of the close relationship between water resources and drought, drought management is an important element of water resource management (Bazza 2002). Two general categories of measures can be applied for the proactive management of water resources for drought preparation and mitigation, both planned in advance, these are long-term and short-term (Cancelliere et al. 2007, Giuseppe Rossi, Vega, and Brunella Bonaccorso 2007). Long-term measures are aimed at improving the reliability of water supply systems to meet future demands under drought conditions through structural and institutional measures. Such measure are increasing water storage capacity, adopt water saving technology and recharging ground water levels. Short-term measures try to face the incoming drought conditions within the existing framework of infrastructures and management policies. These types of measure represent the actions taken during a drought contingency plan, and they are gradually implemented to reflect the progressive onset of the drought. Each of these two types of measures can be further categorised into three sub-categories: water supply oriented, water demand oriented and drought impact minimization (Yevjevich, Cunha, and Vlachos 2012). Supply oriented measures aim at increasing the available water supplies, while demand oriented focus on improving the efficiency of existing water resources. Drought impact minimization measures focus on reducing the economic, environmental and social impact of droughts (G. Rossi, Castiglione, and B. Bonaccorso 2007).

Generally a combination of both long-term and short-term measures is preferable, while the interrelation of supply, demand and minimization efforts need to be taken into

account (Bazza 2002). Still a general solution that encompasses all of these issues is difficult to establish, given the complexity of water resource behaviour and the time and costs involved in implementing efficient long-term measures. One measure that can be easy and cheap to implement, and that is also relevant to the previously described context is the development of an early warning system. This represents a long-term measure for drought impact minimization (FAO 2001; Dziegielewski 2003; G. Rossi, Castiglione, and B. Bonaccorso 2007) that will serve alongside the short-term measure already imposed by the WWTP at De Punt, the downscaling procedure. Such an early warning system could be implemented by making use of the predictive power of Machine Learning (ML) on the streamflow of the river that supplies the plant.

ML is the study of computer algorithms that improve automatically through experience (Mitchell 1997). It is a branch of Artificial Intelligence (AI) that makes use of training data to make predictions, without it being specifically programmed to do so (Koza et al. 1996). On category of ML is called "Supervised Learning" and the purpose is to train a model to predict a given variable (target) given example inputs (predictors). In such a case the outcome is known and the model is trained to learn the rules that map the inputs to the outputs.

Preventing the effect of droughts requires accurate and reliable streamflow predictions, and represents a crucial step in water resource management (Wang et al. 2019; Rezaie-Balf et al. 2019, Gauch and Lin 2020). Predicting streamflow is a spatio-temporal forecast that makes us of past streamflow values and meteorogical data (temperatue, humidity, precipitation, etc.) (Gauch and Lin 2020). The non-linear and non-stationary nature of such processes makes forecasting a difficult challenge, especially at the extreme end of these behaviours, such as droughts and floods. (Meng et al. 2019). ML based models are based on historical observations, they are empirical and easy to implement and do not require knowledge on the underlying physical processes (Liu et al. 2015). Over the years, ML techniques have received significant attention with various models being applied within water sciences with satisfactory results, such as Artificial Neural Networks (ANN) (Wen et al. 2019; D. Zhang, Lindholm, and Ratnaweera 2018; Hussain and Khan 2020; Parisouj, Mohebzadeh, and T. Lee 2020; Tongal and Booij 2018), Support Vector Regressor (SVR) (Meng et al. 2019; Hussain and Khan 2020; Asefa et al. 2006; Parisouj, Mohebzadeh, and T. Lee 2020; Tongal and Booij 2018) and Random Forest (RF) (Tyralis, Papacharalampous, and Langousis 2019)

Recently, Extreme Gradient Boosting (XGB) (T. Chen and Guestrin 2016) has received significant attention within the earth sciences community, outperforming many of the previously used models and proving itself a reliable method for predictions (Fan et al. 2018; Xiao et al. 2018; R. Zhang et al. 2019; X. Chen et al. 2018; Xia et al. 2017, Ni et al. 2020). It is similar to the RF model in that they both rely on the DT as the base estimator, in an effort to combine multiple "weak" learners to obtain a "strong" learner. Where they differ is in the way they are built internally. Where RF is based on parallel ensembling (training multiple "weak" learners and average the prediction amongst all of them), XGB is basted on GB (training each successive "weak" learner to correct the error made by the previous one) to obtain a "strong" learner). Despite its use over a wide domain of fields (X. Chen et al. 2018; Xia et al. 2017), with water resources being one of them (Xiao et al. 2018; R. Zhang et al. 2019), efforts to use it for streamflow modelling

have been minimal (Ni et al. 2020). As such studying the prediction performance of XGB for streamflow forecasting will provide possible applications for it, and also give a better understanding of the model. The capabilities of an efficient and accurate drought early warning system provides water managers with the tool they need for the efficient use of short-term measures for drought impact minimization. The knowledge gained from the model is valuable for the ML field for it is a new technique that has shown good results so far. Also, the reliance on the DT for the "weak" learner makes it is easier to interpret than many of the popular models, a crucial issue when addressing the uncertainty of the model for decision making.

Lasswell and Kaplan describe decision making as "forward looking, formulating alternative courses of action extending into the future, and selecting among alternatives by expectations of how things will turn out" (Lasswell and Kaplan 1950). The predictive capability of ML is appealing for decision makers as it promises not only to determine the outcome of their decisions, but also provide explanations for it that will diminish the uncertainty in that decision (Pielke). Uncertainty refers to the degree of outcome consistency compared to our perceived and understood expectation (Pielke 2001), and one way to reduce it is through model interpretability. Interpretability has no formal definition, however good descriptions are given by Miller (Miller, 2017) "the degree to which a human can understand the cause of a decision" and Kim (Kim, 2016) "the degree to which a human can consistently predict the model's result". The easier it is to understand why a certain decision was made, the more trust there is be in the prediction, therefore the less uncertainty there is in the model (Molnar). Interpretability of a model can be categorized in two ways, global and local interpretability (Molnar; Velez, 2017). Global interpretability gives an understanding on the distribution of the outcome based on the features used, while local interpretability explains a particular prediction instance (or group of instances). Both types of interpretability can be obtained by using certain algorithms to generate them, one such algorithm is known as the Shapley Value (SV).

Borrowed from game theory (Shapley, 1951; Roth, 1988) the SV represents the individual contribution and individual player has in the gain resulted from a cooperative situation amongst multiple players. For regression, the features used are the players and the gain is the targeted predicted outcome, as such we can use SV to determine the individual contribution of each feature used in predicting streamflow. Such knowledge would help quantify the reasoning behind the prediction in a manner that is easy for a decision maker to interpret and trust. This quantification could lead to further insights for the system under investigation as well as future possible hypothesis than need to be tested. Management and operational research have seen significant use of the SV, with applications in cost allocations (Landinez-Lamadrid, 2017) and networking applications (Cesari, 2018), with the water sector seeing some interest (Sadegh, 2011; Jafarzadegan, 2013; Schmidt, 2020). However, to the authors knowledge, no efforts have been made to use the SV for model interpretation in streamflow prediction application.

Such a technique would provide managers with the insight they need in the long-term to carefully plan and execute the short-term measures at his disposal. If a decision maker understands why the model predicted a certain value for the streamflow sometime in the future, then he will have more trust in the model, and in turn will be able to justify the need for a certain decision to be made. In the case of this project, the complicated downscale

procedure of the plant can be efficiently implemented if the decrease in flow is detected and understood in advance. Therefore this project aims at designing a predictive model for the streamflow of the Drentsche Aa river, and obtain global and local interpretations of the models predictions.

1.2 Methodology and Structure

The methodology represents the series of steps (or tasks) taken in the execution of this project. The IBM Foundational Methodology for Data Science is used (Figure 2).

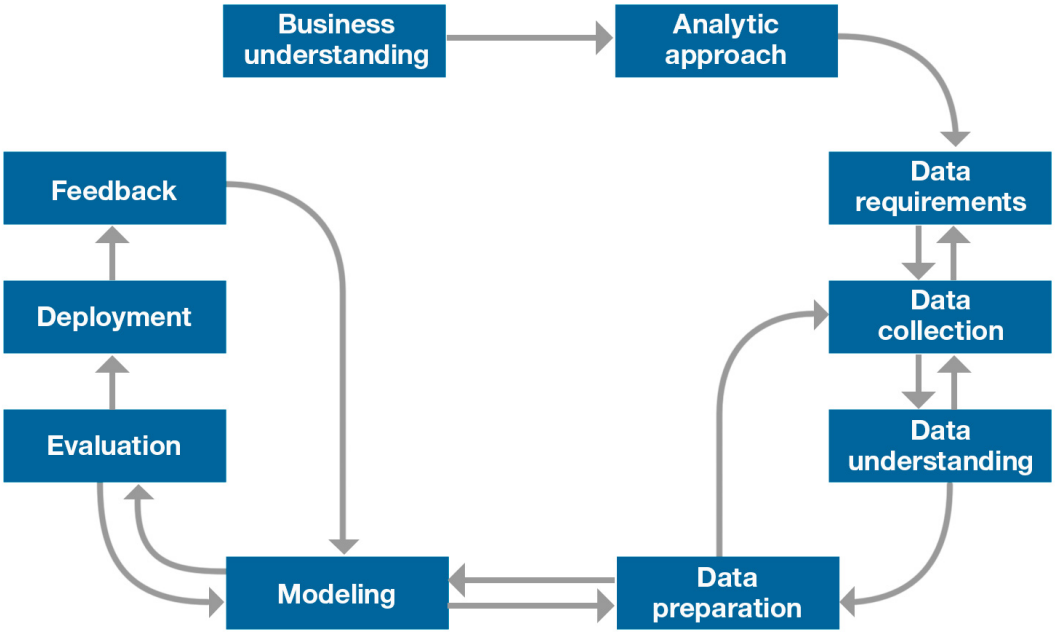

Figure 2: Foundational methodology for data science. (source: IBM "Why we need a methodology for data science")

In *Business Understanding* (BU) (the previous section of this report) the problem to be solved is discussed and explained together with possible solutions that are to be tested. Specifically this step involves discussions with domain experts, involved stakeholders such as members of the WB and of the WCG, and within the design team that help in the better understanding of the process. The result of this step is a description of the process to be modeled and the problem to be solved. For the *Analytic Approach* (AA) step, the problem is expressed in a statistical and ML context and the most suitable solutions are identified and described.

In the next four steps (also referred as *Data Wrangling*) (DW), we define, collect, inspect and prepare the data to be used. The *Data Requirements* (DR) will define the way in which the *Data Collection* (DC) task will gather data from the all of available sources. *Data Understanding* (DU) represents the use of descriptive statistics and visualization techniques to asses the data content, quality and discover initials insights about it. For the *Data Preparation* (DP) task, the methods and steps taken to clean, combine and transform data are explained

In the *Modeling* (MLG) step, ML methods are used in order to obtain the model that will predict the river flow. The phases of this step are *Feature Selection* (FS), where the most optimal combination of features is determined, and *Model Tunning* (MT), where the model with the best performance given a data set is determined. The result of this step is a combination of data and trained models that are to be evaluated.

In *Evaluation* (EV) the data and models identified in the modeling step will be evaluated against a series of performance metrics in order to understand the quality and efficiency of the given models. The result of this step will be the best performing model and the performance criteria by which it was selected.

Finally, in *Deployment* (DEP), the optimal model is trained on the entire dataset after which a series of Model Interpretation (MI) techniques will be deployed on that model in order to explain the model predictions. The result of this step will be global and local explanations of the model predictions and errors and their relationship with the selected features.

The structure of this document is as follows. Section 2 corresponds to the *Business Understanding* step, Section 3 to the *Analytic Approach*, Section 4 contains all data related functions, Section 5 contains *Modeling* and *Evaluation*, and Section 6 the *Deployment* and *Feedback* tasks.

2 Business Understanding

In this section the problem is expressed from a business perspective, also the company goals and stakeholders are presented together with the research questions the project will focus on.

2.1 System Description

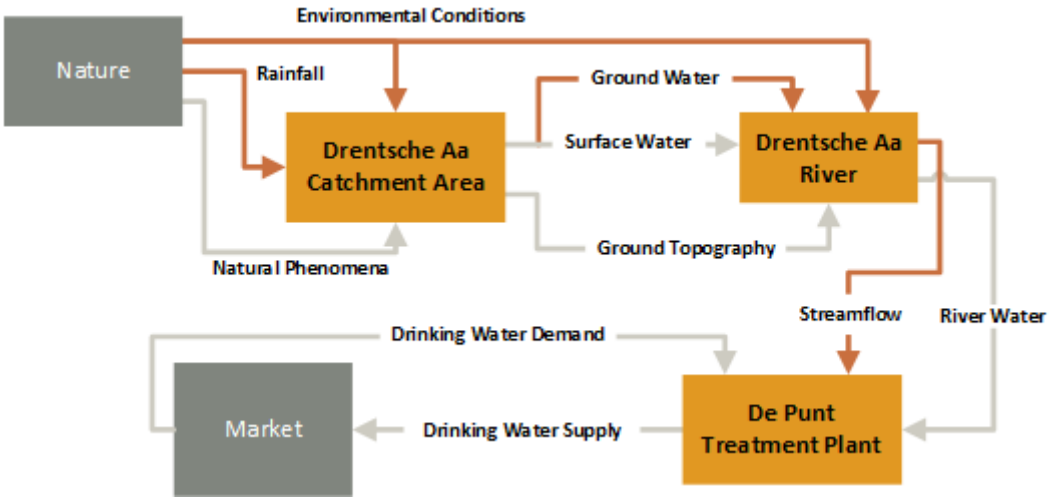

Figure 3: Description of the system and its elements. Elements within the scope are depicted in orange and elements outside the scope in gray.

Nature: provides the *Environmental conditions* (temperature, humidity etc.) that act as controls for the catchment and river representing the environment under which the two systems behave. Besides this, the main input to the catchment, the *Rainfall*, and the *Natural Phenomena* (Evaporation, Infiltration etc.) are provided by this component.

Drentsche Aa Catchment Area: represents the environment surrounding the river. The *Water Balance* of the area varies by means of various *Natural Phenomena* and is dictated by specific *Environmental Conditions* and amount of *Rainfall*.

Drentsche Aa River: can be regarded as the central component of this study, as the *Streamflow* is the central characteristic to be modeled. Its behaviour is ensured by the *Ground Topography* under which it flows and is dictated by the *Ground Water* levels as well as the *Surface Water* of the catchment area.

De Punt Treatment Plant: is the main producer of drinking water. Under the *Management* of the WCG and based on the *Water Demand* from Consumers together with *Streamflow* from the river, they provide the *Water Supply* for Consumers using the *River Water*.

2.2 Problem Description

2.2.1 Drought

Water treatment plants in general work by processing infeed water from a given source using various techniques. As such the availability of this infeed is crucial for the proper functioning of the plant. Lack of such infeed can lead to various problems within the treatment process.

Drought is defined as a period of time which experiences bellow-normal precipitation, and is typically accompanied by high temperatures. It is difficult to pinpoint the time interval in which a drought occurs, as unlike sudden events, like hurricanes, it takes time for the effects of bellow-average rainfall to take their effects.

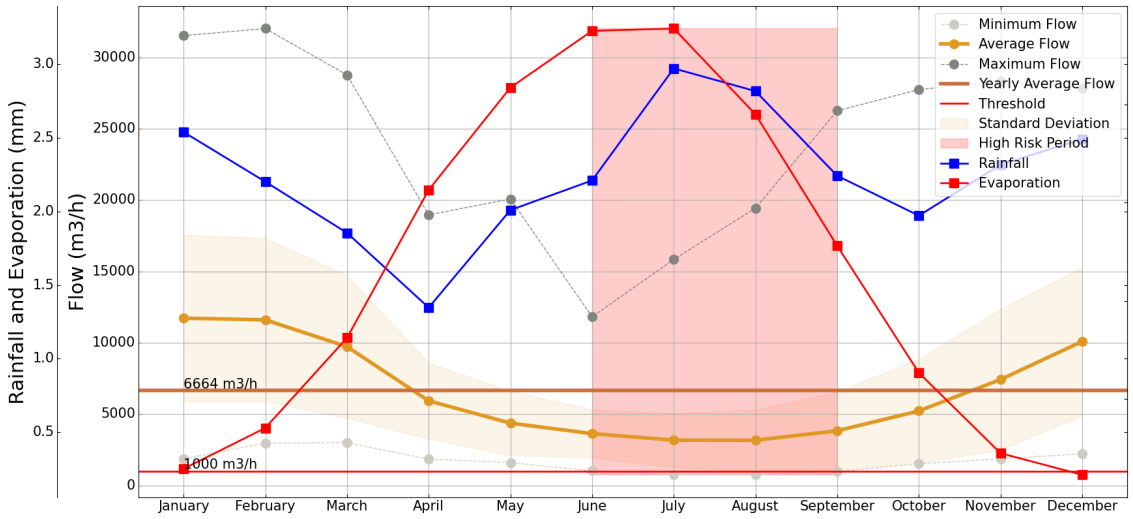

Figure 4: Monthly average values for flow (orange), precipitation (blue) and evaporation (red)

The flow values of the Drentsche Aa river are depicted in Figure 4. Following the rainfall and evaporation evolution throughout the year, we notice that during the winter and autumn months, low amounts of evaporation and high amounts of rainfall generally translate to higher flow values. In the summer the two phenomena balance each other out, with the flow remaining consistently low.

The low level of the river during the summer raises issues as far as the treatment plant at De Punt is concerned. It was designed to output a nominal flow of drinking water of 840 m^3/h, and the typical amount it needs in the infeed for this is around 850 m^3/h. On top of this there is the amount of water the plant needs to keep downstream of the intake point of 100 m^3/h such as not to deplete the entire river. As such, periods of the year in which this flow might fall bellow the minimum threshold represents a period of high risk for the plant (red area).

2.2.2 Affected Components

Should a bellow-threshold event occur, the plants output and performance will be scaled down, while bypassing water from the nearby ground water treatment plant to maintain the plant functioning. However, ground water is much colder than the surface water of the river, and the bacteria used in the process are highly sensitive to variations in temperature. Oxygen dissolves in water at different rates depending on temperature, low temperatures result in a high amount of dissolved oxygen, while the opposite is true for high temperatures. As such the process of scaling the output of the plant needs to be slow enough as not to disturb the balance of the processes in the plant. The main components affected by this are the Active Carbon Filtration (ACF) and the Slow Sand Filtration (SSF) steps (Figure 5)

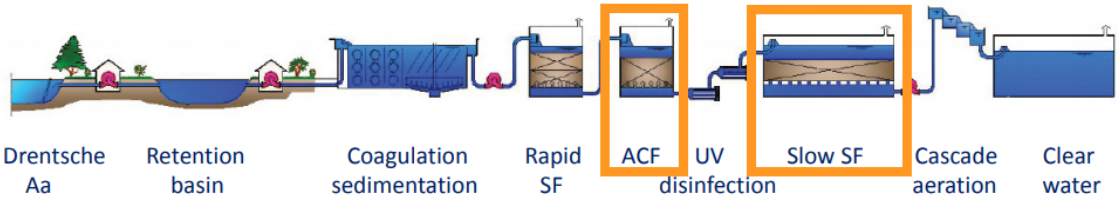

Figure 5: Process components affected by insufficient water in the system

- **Active Carbon Filtration:** In this step, activated carbon is used to remove dissolved substances through adsorption, also the porous structure of active carbon makes it an ideal carrier of biomass.
- **Slow Sand Filters:** In this process, a complex biological film, present at the top of the sand is used to treat the water.

Damage to any of these two processes could lead to extremely high cost (replacement, repair) and prolonged periods of downtime (repopulating bacteria, cleaning filters and coagulation).

2.3 Stakeholder Identification

Stakeholders are key individuals or entities involved or affected by a project. They have different levels of interest and influence in the project and aligning their goals and objective is crucial for the succesfull implementation of any project. In this section the key stakeholders of this project are presented and described (Figure 6).

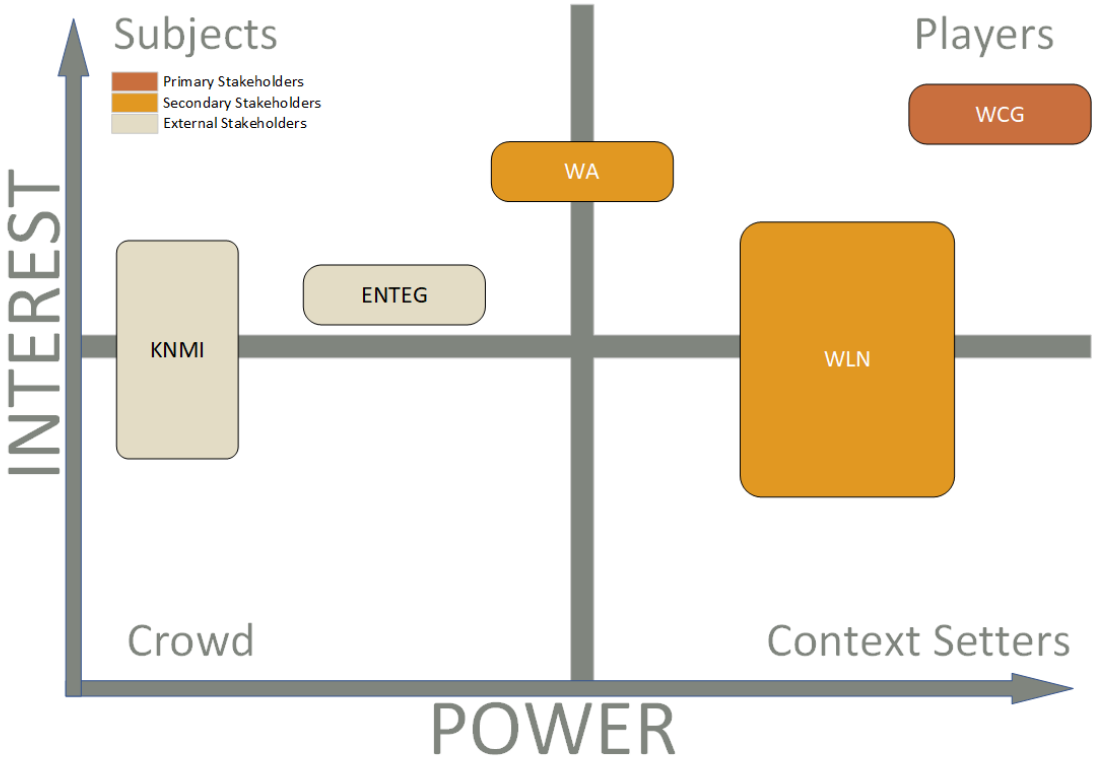

Figure 6: Stakeholder identification diagram depicting primary stakeholders/problem owner (dark orange), secondary/key stakeholders (light orange) and external stakeholders (gray) according to their power/interest

- **Water Company Groningen (WCG):** As the owner of the plant, they represent the primary stakeholder and the problem owner. The correct functioning of the plant is their responsibility and therefore they have the highest influence and interest in this project.
- **Water Authority Hunze and Aa's (WB):** The duties of the water authorities according to the Water Act 2009, are the development of of a management plan for the water system and setting up general rules and regulation for all activities in and around water, including floods and draughts. As such they have a significant interest in the outcome of this project.
- **Water Laboratory North (WLN):** A subsidiary of the WCG, the goals and objectives of WLN is to act as consultant in all matters regarding water technology, perform water analysis and execute research on various drinking water related subjects. As the entity responsible with executing the project, the influence of this stakeholder is relatively high, while it's interest varies from moderately high to moderately low, seeing as not all of its objectives align with this project.
- **Royal Netherlands Meteorological Institute (KNMI):** Acting as the Dutch national weather service, the KNMI tasks are weather forecasting and monitoring

of weather, climate, air quality and seismic activity. They are the main source of meteorological data, and as such they might display some interest in this topic.

- **ENgineering and TEchnology institute Groningen (ENTEG):** Having the task to analyse, explore and design technologies that integrate different engineering sciences, it focuses on various activities in the processing and production sector. The methods and techniques used in this project might represent a key interest for them.

2.3.1 Problem Statement

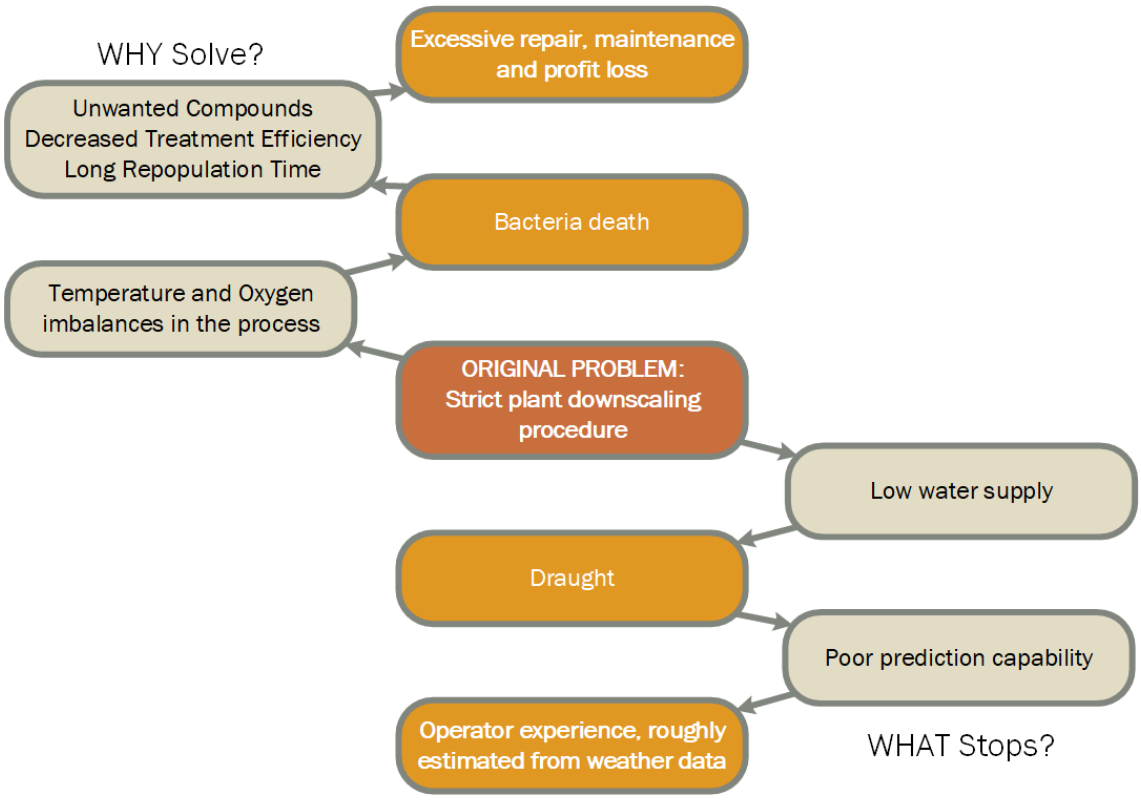

Figure 7: Why-What analysis depicting the original problem (dark orange), problems caused at different levels (light orange), difficulties in solving the original problem (right gray) and reasons to solve the original problem (left gray)

Figure 7 displays a Why-What analysis, describing the main causes of the problem, what are the factors that stop this problem from being solved and the main reasons why the problem should be addressed. Periods of draught are hard to forecast, they take place over and extensive period of time and can be affected by various factors. Failure to predict a shortage in river water caused by this draught could lead insufficient water supply in the plant intake. To combat this the plant must be dowscaled to match the supply of the river, however, the way in which the plant's capacity is reduced need to follow a strict schedule. Since bacteria present in the process are highly sensitive to temperature and

oxygen, disturbing this balance could lead to the death of these organisms. Loss of biomass in the filters result in a decrease in water quality caused by the inefficiency of the treatment process, unwanted compounds resulted from dead bacteria and a long repopulation time required to bring the process back to normal operating conditions. These result in excessive repair and maintenance costs to clean the unwanted compounds and repopulate bacteria as well profit loss caused by decreased water quality. As such the problem statement of this project can be formulated as follows:

> *The inability to accurately predict periods of draught and their effect on the supply of water in the Drentsce Aa river combined with a strict downscaling procedure of the plant may lead to Temperature and Oxygen imbalances in the process that could cause the death of the bacteria present in the filters resulting in excessive repair and maintenance costs caused by the long repopulation time of the bacteria and also profit loss from discharging water during repopulation.*

To solve this issue, a predictive model will be developed that will be able to predict the supply of the river within a reasonable time period, and provide an explanation of the prediction based on input parameters. The lead time of the forecast will be determined from model performance and scaling speed of the plant. This model will allow for the safe and efficient scaling down of the plant to occur, as well as increase our understanding on the conditions under which drought occurs.

2.4 Goals and Research Questions

Organizations establish goals in order to determine what needs to be done to achieve their targets. A goal is a statement of intention aimed at informing members of the organization, but also involved stakeholders on the issue that needs to be addressed. They are broad, and are part of the big picture view that needs to be narrowed down further as the problem and its detailed aspects become more clear. As such the goal of this project is stated as follows:

> *Develop a Decision Support Tool to predict the flow of the Drentsche Aa river and assist the operators of De Punt Water Treatment Station safely scale down the plant in order to avoid a possible shortage of water or process imbalances in the system during periods of drought.*

To understand how and if this goal can be achieved, research question will be developed next. Research questions are essential elements in both quantitative and qualitative research. They represent the knowledge that a certain research project aims to uncover, and offer a starting point in the execution of a design project. For this project, the knowledge required is twofold. First the physical aspects of the situation, plant and river, need to be properly understood. Limitations of the plant compared to the river evolution, and the extend to which plant can safely follow the river evolution (specifically drop in flow) are essential in determining whether a given solution can realistically solve the problem at hand. Second, the contribution of ML needs to be underlined. The predictive performance of the selected models as well as the information (features) required to make the predictions will provide the ML and Water Management fields with a deeper understanding into

the possibilities and application of these models. Besides this, the explanations given by the model on the predictions will provide managers and operators with the empirical reasoning required to make informed decisions. As such the research questions of this project are stated as follows:

1. *How does the evolution of the river compare to the scaling possibilities of the plant?*
    - *What are the plant dynamic limitations?*
    - *What are the worst case scenarios of the river dynamics according to plant limitations?*
    - *What are the risks of the river's worst case scenarios?*
2. *How can Machine Learning be used to predict and explain the flow of the Drentsche Aa river?*
    - *What is the model that provides the best performance?*
    - *What are the most important features used by the model when making predictions?*
    - *What is the global effect of these features on the predicted value of the river flow?*

2.5 Objectives

Objectives represent the steps taken in order to achieve the overall goal and to answer the previously raised research questions. They represent the individual milestones used to track the execution of the project. Developing objectives in a S.M.A.R.T. (**S**pecific, **M**easurable, **A**ttainable, **R**elevant, **T**ime-bound) way makes sure that we set realistic expectations from the project and avoid any situation in which poor planing leads to failure in their achievement. The specificity of these objectives is given by the previous description of the problem, plant and river dynamics, model error constraints as well as desired target to predict. Measurability is assured by the choice of performance measure selected to validate the success of the models prediction. Given the vast range of similar successful attempts at predicting the same (or similar) target variable (streamflow), the objectives are achievable. The high costs incurred in case of failure to predict a river worst case scenario, as well as the vast amount of literature developed on this topic ensures the topic of this project remains relevant. The time in which this project needs to be completed is 5 months, also given the fact that a typical model (given a reasonably sized dataset) takes anywhere between 1 to 20 seconds to train, the time imposed for the completion of the project is reasonable. As such, the objectives of the project are as follows:

1. *Train and test a ML model capable to predict the streamflow values of the Drentsche Aa river in a way that provides low risk when predicting extreme cases*
    - *Select the 10 best features that give the best performance.*
    - *From the 10 best features select the subset that minimizes the MSE*
    - *Using the best subset, tune each individual model to minimize the MSE.*

2. *Select the best performing model and use the Shapley Value (SV) technique to explain it's predictions.*
    - *Compute the SVs for the best performing model*
    - *Describe the model global characteristics of the features using their SVs*
    - *Describe the model local characteristics of the features during periods of drought using their SVs*
3. *Develop a software solution able to maintain a up-to-date database of the inputs required to make the predictions and provides the interpretation data for the model predictions*

3 Analytic Approach

In this section the plant constraints and detailed aspects of the problem are given, the methods by which the problem will be handled are presented and the objectives of the project are formulated.

3.1 Downscale Procedure

To begin to understand the problem the plant faces, the plant downscaling procedure in case of insufficient supply of water is described (Figure 8)

Surface Water Temperature (SW)	25.0 °C
Ground Water Temperature (GW)	11.0 °C

SW (m^3/h)	ΔSW (m^3/h)	GW (m^3/h)	Total Flow (m^3/h)	Temperature (°C)	Δ T (°C/day)	Production		Day
						SW (%)	GW (%)	
850	-	0	850	25.0	-	100	0	0
790	-60	60	850	24.0	-1.0	93	7	1
730	-60	120	850	23.0	-1.0	86	14	2
670	-60	180	850	22.0	-1.0	79	21	3
610	-60	240	850	21.0	-1.0	72	28	4
550	-60	300	850	20.1	-0.9	65	35	5
405	-145	300	705	19.0	-1.1	48	52	6
300	-105	300	600	18.0	-1.0	35	65	7
225	-75	300	525	17.0	-1.0	26	74	8
165	-60	300	465	16.0	-1.0	19	81	9
115	-50	300	415	14.9	-1.1	14	86	10
70	-45	300	370	13.6	-1.3	8	92	11
30	-40	300	330	12.3	-1.3	4	96	12
0	-30	300	300	11.0	-1.3	0	100	13

Figure 8: Plant downscale procedure (Source: WLN))

Should the river run out of water, the plant needs to downscale production in order to avoid lack of water in the system. To do this, the amount of surface water in the intake (SW) is gradually reduced while compensating with water from the ground water plant (GW) up to a maximum of 300 m3/h. After this, the intake from GW is kept at 300 m3/h while intake from river water is reduced. However, GW temperature is much lower than that of river water, as such the temperature in the system will begin to drop as GW is feed in. As the amount of SW is reduced, the amount of oxygen available to the bacteria will also drop. To compensate for this, the GW supplement is saturated with oxygen, still if the water intake is reduced to fast this compensation will not be enough to ensure proper living conditions for the bacteria. The most important aspect of this downscaling is the ammount by which the temperature drops with each day (delta T). In order to avoid severe imbalances in the system, the rate of change of temperature need to be maintained

around -1.0 °C. This is done because of the fact that oxygen dissolves much slower at high temperature, while the bacteria oxygen uptake rate remains constant. At low temperature however, oxygen dissolves much faster, as such after reaching the 300 m3/h of GW intake, the speed at which the SW can be reduced is increased.

From this procedure we see that the maximum prediction time for the plant is 12 days, as that is the time in which the plant can go from full capacity to a complete stop. Another possible prediction interval is 6 days, as after that the plant can downscale much faster.

3.2 River Dynamics

The next question is whether the river dynamics (drops and increases in flow), allow for a safe downscale. Next we will analyse how realistic and safe this procedure is compared to the statistical evolution of the river flow.

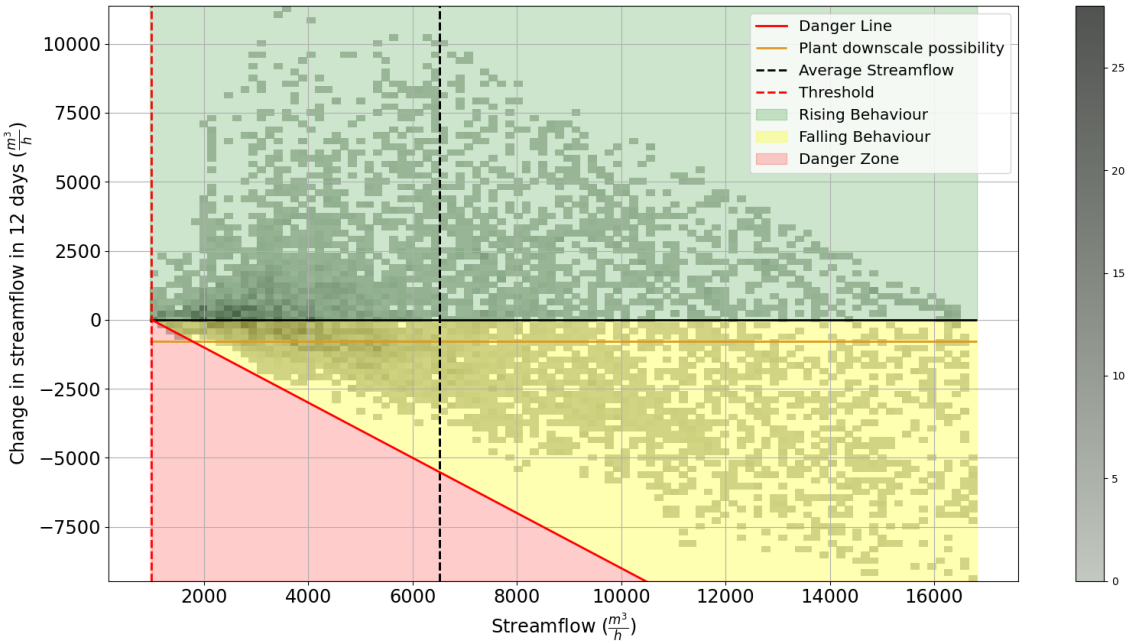

Figure 9: Bivariate distribution of the river dynamics (flow and change in flow) over a period of 12 days compared to the plant downscaling possibility (from full capacity to no production) and imposed flow threshold

Figure 9 depicts the historical change in the flow of the river over 12 days. On the x axis we have the flow distribution of the river, while on the y axis we have the distribution of the river flow change over 12 days. From the figure we can see that on average the flow of the river is about 6500 m^3/h. The negative relationship between streamflow and change in streamflow indicate a limiting effect at the extreme ends of the flow interval. This simply means that is is highly unlikely for the flow to increase any further from already high values (river reaching maximum capacity), and vice-versa it is highly unlikely for the flow to decrease any further at low values (no more water for the river to lose).

Comparing the river dynamics with the plant limitations is done by displaying the plant dowscaling possibility in 12 days, time in which the plant can go from full capacity to no river intake (orange line). The significance of samples bellow this line is that of a river flow drop that occurs faster than the plant can downscale. Such a scenario means that whatever the plant does, it will never be able to follow the river flow, and runs at a high risk of insufficient water supply. However at high enough values, despite the fact that the river flow drops faster than the plant can downscale, the amount of water already present in the river means that there will still be plenty of water for the plant to safely resume operations. A dangerous situation (red diagonal line) is that of a river drop that will result in a river flow bellow the imposed threshold (red vertical dotted line). Samples bellow this line represent scenarios in which the river drop is so severe that the plant will have to begin downscaling, otherwise it will certainly run out of water in the intake.

The highest risk for the plant, are those samples that fall bellow both the danger line(red diagonal line), and the plant downscale possibility(orange line). Sample bellow these lines represent scenarios in which the flow of the river will drop bellow the imposed threshold, and will do so faster than the plant can downscale. This is the most dangerous situation, and it is the main scenario the problem owner would like to avoid. From Figure 9 we can see that this situation is highly unlikely to occur, as the historical data suggest that the flow cannot fall so drastically as to outrun the downscale possibility of the plant. Any scenarios in which the flow could fall bellow the imposed threshold can be safely covered by the existent procedure.

3.3 Methods and Tools

In this sections the mathematical and statistical tools and algorithms that are to be used in the analysis and design of the predictive model are described. Initial considerations on the predictive task are given first followed by descriptions of the individual algorithms used to predict the required target, followed by definitions of the performance indicators that asses the quality of the predictions and finally the Shapley value is explained.

3.3.1 Machine Learning Prediction

Predicting is the task of computing the value of a target variable as close as possible using a set of input variables. The function of a predictive model is to map the values of the inputs used to the values of the desired outcome. The ML task is to determine the optimal input matrix X and predictive model f that correctly give the values of the target variable y. However, since no prediction or forecast is completely accurate, there is an amount of irreducible error e than need to be taken into account. This error is the result of noise in the available data, incompleteness of said data and complexity of the processes to be modeled. Mathematically this can be expressed as:

$$y = f(X) + e(X), \tag{1}$$

where y-target outcome, X-input matrix, f-model, e-error.

The design approach taken in this project will differ to that of mainstream streamflow prediction. Instead of attempting to forecast the actual value of the streamflow, we will predict the change in streamflow that will occur within the required time according to the

downscale strategy. Doing so, will allow us to compute Shapley values that will interpret and describe what causes the change in river streamflow from the current value, instead of describing the absolute value the streamflow will have in the future. This represents a more dynamic approach to modeling the amount of water in the river by focusing more on the difference (or derivative in the continuous case) of the streamflow from one timestamp to another. This can be expressed as following:

$$\frac{\Delta y}{\Delta t} = y(t + \Delta t) - y(t), \tag{2}$$

where: $\Delta y/\Delta t$-change in flow in Δt days, y-streamflow, t-current timestamp, Δt-prediction time interval. Given this, the target of the model is

$$\frac{\Delta y}{\Delta t} = f(X) + e(X). \tag{3}$$

Replacing this in equation 2 we can see how the final predicted value of the target is obtained:

$$y(t + \Delta t) = y(t) + f(X) + e(X), \tag{4}$$

As such the model will predict the change in stream flow that will occur in the next Δt days.

3.3.2 Sample Weights

Another design aspect taken into account, is the emphasis placed on low streamflow/fast decreasing streamflow samples (Figure 10). The left side of the figure depicts the sample weights for flow change (top) and flow (bottom), while the rights side represents the result of the sum of these weights. Doing so, will result in higher weights for cases in which the flow of the river was low to begin with and that flow decreased severely in the future. The worst case for the plant is a scenario in which there is little water to go by and in the near future that supply will decrease drastically, to such a degree that there will be insufficient water in the infeed. As such the modeling focus will be placed on low flow values that exhibit a rapid decrease in the near future. The less water in the river, and the faster that streamflow will decrease in the future, the more accurate the models will have to predict it.

Using these sample weights we can identify samples that can be considered to be in a period of drought by sampling a certain percentage of the top highest weights. In Figure 10, the top 5% of sample ordered by descending order of sample weight are shown in orange, these represent historical samples in which the flow had a low value and/or it was about to decrease significantly in the next six days, twelve day weights are obtained in a similar manner. Analysing these instances will provide information on how the river streamflow behaves during periods of drought.

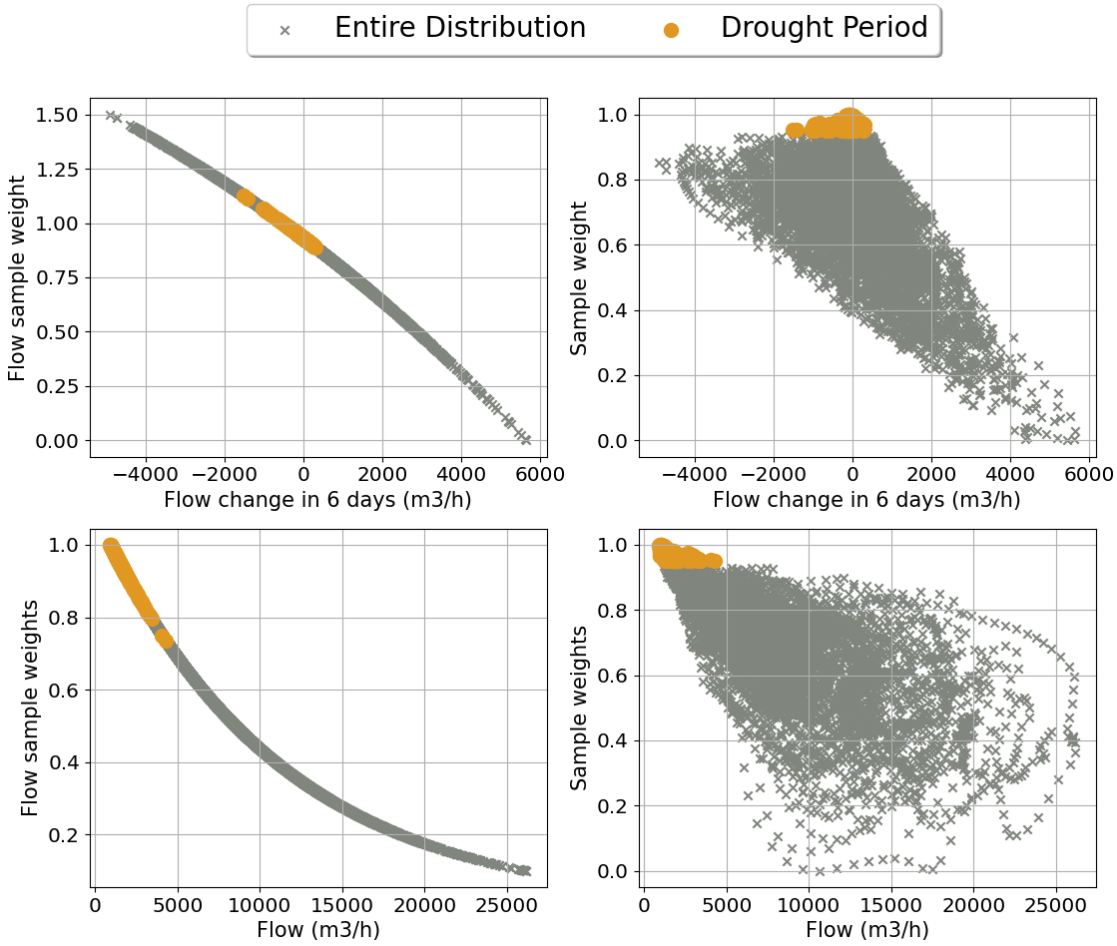

Figure 10: Sample weights depending on flow and future change in flow

3.3.3 Machine Learning Models

Decision Trees

A decision tree is a decision support tool that performs tree-structured decision tests in a divide-and-conquer way. The tree is comprised of two elements, nodes and leafs. Each node of the tree is associated with a feature test, that performs a binary split on the samples used for the test [23]. Binary splits Each leaf node represent a particular label assigned to the samples in that leaf. Leaf nodes represent the final prediction $\hat{y}_i = f(x_i)$. The most important concept of building decision trees is the process of splitting the data at each node. The learning algorithm is a recursive process, where at each step (node), the given data is split into subsets. Each of these subsets are then used in the next step of the process and are further split into more subsets. To perform a split, a feature X_i is selected from the columns of X, as well as the target y. Samples x_j are ordered in increasing order and the mid-point value α is chosen between each neighbouring pair of ordered samples to split the samples at that threshold such that

$$\begin{aligned} X_i^L &= \{x_{ij}; x_{ij} < \alpha \\ X_i^R &= \{x_{ij}; x_{ij} \geq \alpha \end{aligned} \tag{5}$$

and similarly for y

$$\begin{aligned} y^L &= \{y_j; x_{ij} < \alpha \\ y^R &= \{y_j; x_{ij} \geq \alpha \end{aligned} \tag{6}$$

After the split, the mean value of the samples of y belonging to each split is taken as the output prediction $\hat{y}$ such that:

$$\begin{aligned} \mu^N = \frac{1}{n} \sum_{j \in N} y_j = \overline{y}^N \\ \hat{y}^N = \overline{y}^N \end{aligned} \tag{7}$$

where N is the node at which the split was made, n is the number of samples in that node and $\overline{Y}^N$ is the mean values of the samples in node N.

At each node the error is computed as:

$$MSE^N = \frac{1}{n} \sum_{j \in N} (y_j - \hat{y}^N)^2 = \frac{1}{n} \sum_{j \in N} (y_j - \mu^N)^2 \tag{8}$$

The error at the two nodes created by the split is:

$$\begin{aligned} MSE^L = \frac{1}{l} \sum_{j \in L} (y_j - \mu^L)^2 \\ MSE^R = \frac{1}{r} \sum_{j \in R} (y_j - \mu^R)^2 \end{aligned} \tag{9}$$

The idea behind building a decision tree is to select a threshold value for feature Xi that minimizes the total error after the split, $MSE^L + MSE^R$. The next step in the algorithm is to perform the same splitting procedure for all the remaining features of X, and select the feature and threshold that minimizes the error overall. The general assumption is that the error in the node is higher that the error after the split such that:

$$E_1 \geq E_2 \tag{10}$$

where:

$$\begin{aligned} E_1 = MSE^N \\ E_2 = MSE^L + MSE^R \end{aligned} \tag{11}$$

Using this we can explain the decrease in error, or the information gain, caused by the split on feature i and feature values = $alpha$ to be:

$$G_\alpha^N = MSE^N - (MSE^L + MSE^R) \tag{12}$$

The importance of the node where the split was made is given by:

$$I_N = n \cdot MSE^N - (l \cdot MSE^L + r \cdot MSE^R) \tag{13}$$

where n - number of samples arriving at node N, l - number of samples ending up in the left child of the node, r - the number of samples ending up in the right child of the node. The importance of feature X_i is the fraction of the total node importances were feature X_i was used to perform a split:

$$FI(X_i) = \frac{\sum I^J}{\sum I^k} \tag{14}$$

where j - nodes were feature Xi was used to perform a split, k - all nodes of the tree.

Random Forest

Random Forests (RF) are ensemble methods used for classification and regression by constructing multiple decision trees using a random subset of features from the entire feature subset. First created by Tim Kam Ho using the random subspace method and extended to combine the idea of bagging and the random selection of features. The main technique of the RF is the principle of bagging, sample the training set with replacement and combine the predictions of multiple learners in order to achieve a final prediction that is better than that of an individual learner. However, RF differs from typical bagging methods in that it also applies the principle to the features. Each time the RF performs a split on the samples, instead of looking at the whole set of features, the algorithm selects a random subset of features, and performs the split on the feature with the minimal error.

Extreme Gradient Boosting

Proposed by T. Chen and Guestrin 2016, XGB is a scalable machine learning algorithm for boosting trees, that is trained in an additive fashion until certain stopping criteria are met. The predicted value has the following form

$$\hat{y}_i = \sum_{k=1}^{K} f_k(X_i), f_k \in F, \tag{15}$$

where $\hat{y}_i$ is the predicted value, f_k is a decision tree, x_i is the input vector, K is the number of decision trees and F is the space of all possible f_ks. The algorithm minimized the following objective:

$$\begin{aligned} L(\phi) &= \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \\ \Omega(f) &= \gamma T + \frac{1}{2}\lambda\omega^2 \end{aligned} \tag{16}$$

where l is the loss between observed and predicted value, Ω is the regularization term to avoid oerfitting, γ is the complexity of each leaf, T is the number of leafs in each decision

tree, λ is the trade-off to scale the penalty, ω is the vector of leaves scores. To build the model in an additive manner, let $\hat{y}_i^t$ be the prediction of the i-th instance at the t-th iteration and build a new decision tree f_t to minimize the following objective:

$$L^t = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{t-1} + f_t(x_i)) + \Omega(f_t) \tag{17}$$

. Using this method a, decision trees a build succesively to correct any errors made by the trees before them, more details on this procedure can be found in T. Chen and Guestrin 2016.

3.3.4 Performance Measures

To understand how good a particular model is at predicting new samples, various performance measure are used that help identify the goodness-of-prediction of the model. Using these values, we will get an idea on how the model is behaving, also based on this metrics the best model will be chosen. The measures used in this project are Mean Absolute Error (MAE), Mean Squared Error (MSE) and Coefficient of Determination (R^2).

Mean Absolute Error

The MAE is the arithmetic average of the absolute error, where the error is the difference between the predicted values and the observed value. It can be expressed as following

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i|, \tag{18}$$

where $\hat{y}_i$ is the predicted value and y_i is the observed value. The MAE is a scale-dependent measure, commonly used in forecasting and timeseries analysis Hyndman and Koehler 2006 and it is one ofe the simplest measures available.

Mean Squared Error

The MSE is the arithmetic average of the squared error, that is the squared of the difference between predicted and observed values and it has the following expresion:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \tag{19}$$

, where $\hat{y}_i$ is the predicted values and y_i is the observed values. The MSE is a second moment measure, and it incorporates the stimators bias (distance from the true value) and variance (distance between between samples). Squaring the prediction error ensures that greater errors are penalized severly in order to ensure a slightly better fit than the MAE.

Coefficient of Determination

The coefficient of determination, or R^2 is the proportion of the variance in the dependent variable that is predictable from the independent variable and it is expressed as follows:

$$\begin{aligned} R^2 &= 1 - \frac{SS_{res}}{SS_{tot}} \\ SS_{res} &= \sum_i (y_i - f_i)^2 \\ SS_{tot} &= \sum_i (y_i - \bar{y})^2, \end{aligned} \tag{20}$$

where y_i is the observed value, f_i is the predicted value, $\bar{y}$ is the mean of the observed data, SS_{res} is the sum of squared residuals and SS_{tot} is the total sum of squares.

3.3.5 Shapley Values

Developed by Lloyd Shapley, the Shapley Value (SV) is a solution in cooperative game theory that assigns a unique distribution (among the players) of a total surplus generated by the coalition of all players (Shapley 1951). In a cooperative situation, in which all involved players contribute and obtain some overall gain, the SV shows how important each player was in the cooperation and what payoff should they receive from the gains resulted from it (Roth 1988). The SV is expressed as follows:

$$\phi_i(v) = \sum_{S \subseteq N - \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S)), \tag{21}$$

where $\phi_i(v)$ is the SV of player i, n is the total number of players, S is the coalition of players, N is the set of all players and v is the values or gain of coalition S.

When referring to a prediction task, the players become the features of the input space and the value of a coalition becomes the predicted outcome of the model. As such, the SV for prediction will describe what effect each feature has on the final value of the prediction. The problem with equation 21, is that the complexity of the computation increases exponentially with the number of players. Thus computing the real values of each players contribution becomes problematic, because one would have to take into account all possible combinations of players.. Recently it has been shown that Decision Trees and tree based model pose an advantage when it comes to computing the real values of the shapley contribution. By following each decision path of a tree, the inclusion of a feature in one such coalition can be determined. As the tree makes binary decision in each node based on one certain feature, the effect of that feature on the final predicted value of the model can be determined by taking into account the predicted values of the leaves corresponding to each individual path. Doing so will provide exact SVs for tree based models, more detailed information can be found in Lundberg, Erion, and S.-I. Lee 2019.

4 Data Wrangling

DW is a term referring to the process of extracting, transforming data into a more usable format, mapping and storing it into a data sink for further use. In this section, all the steps taken in obtaining a clean dataset ready for use in the modeling phase are explained. One thing to note is the choice of nomenclature used in the this stage of the project. The term "signal" refers to the environmental condition being measured (temperature, pressure etc.). The term "feature" refers to the assignment of a location, where the measurement was taken, to each signal. For example attaching the Eelde code (280) to the average temperature signal (TG) we obtain the average temperature at Eelde feature (TG_280). After the Data Wrangling section we will only reffer to features until the end of the report.

4.1 Data Requirements

Data for this project is obtained from two sources. First the meteorological data from Eelde and Hoogeveen stations (station codes 280 and 279 respectively) is fetched from KNMI through the "knmi-py" python package. The data has a daily frequency and all signals are numeric. More detail on the signals can be found in figure 19 of the annex. The second data source is the SCADA system maintained by the WB throughout the rivers catchment area. The data has an hourly frequency and the format is numeric. This data represents the river height measurements from which the river flow will be computed. The unit in which the height is represented is mNAP, and represents the vertical height compared to the national reference point (meters from Normaal Amsterdams Peil).

4.2 Data Collection

The raw data is represented by the river level data, and weather data from the two stations. First the hourly level data is converted into daily data by averaging all values from the same day. The variation of the river level is small on a daily basis so averaging these values is a realistic representation of the daily evolution of the river. Next, weather data from the two station is combined into a single dataset, with each stations code assigned to its respective signal, representing all of the weather data available. Finally the daily level values and weather data are combined into a single data set, and unused feature (time at which that specific signal was measured and also visibility) are dropped.

This final dataset called "Assembled Data" represents the output of this step of the DW task, and will be fed into the DP for further processing.

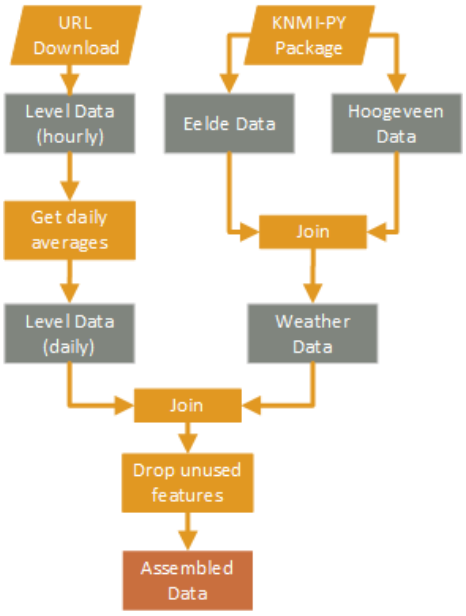

Figure 11: Data Collection steps

4.3 Data Preparation

In this section the steps taken in obtaining a
clean dataset ready for model training are explained. First the data is preprocessed to ensure the validity of the features distributions and correction of any faulty values. Next new features are generated based on the original features in an effort to capture more complex dynamics. For example, using the historical measurements for rainfall we can compute the total amount of rainfall that has occurred in the six days. Doing so will give us an indication of the amount of water that has been introduced in the system in that period of time.

4.3.1 Preprocessing

The preprocessing step is one of the most important steps in any data science project. It is aimed at correct aspects regarding the quality of the data, like out-of-range values, impossible value combinations or missing values.

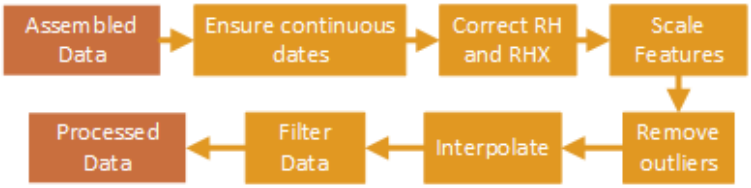

Figure 12: Data Preprocessing Steps

First, missing samples (or dates) caused by sensor errors of computation faults are added to the dataset to ensure a continuous time interval from the first sample to the last one. This is an important aspect of the data because of the time-series nature of the problem. Having the wrong value at a given timestamp could result on prediction error due to incorrect data. Next the rainfall (RH) and maximum rainfall (RHX) are corrected. This is done due to the fact that the sensors that are used to measure this variable return a value of -1 if no rain occurred that day and also the minimum ammount of rainfall they are able to detect is 0.5mm. As such values of -1mm are replaced with 0.1mm, and values of 0mm are replaced with 0.5mm. Some of the signals require rescaling, since for ease of storage they are converted to integer numbers from one decimal floating points, as such they will be rescaled with a scaling factor of 0.1. Following this, outlier values are removed by replacing values that are more than 3 standard deviations away from the mean with an empty value (NaN). After this, the data is linearly interpolated to replace any missing values and outliers that were removed. Finaly, noise in the data is filtered. This is done using an simple moving average filter with a centered window selected such that there is a maximum of 10% correlation loss on the resulting signal.

4.3.2 Feature Engineering

FE is the process of generating new features in the original dataset with the aim at improving the efficiency of the model. Using domain knowledge and data insights, these

new independent variables can be used to improve model accuracy, reduce training times and provide insights into the process being modeled.

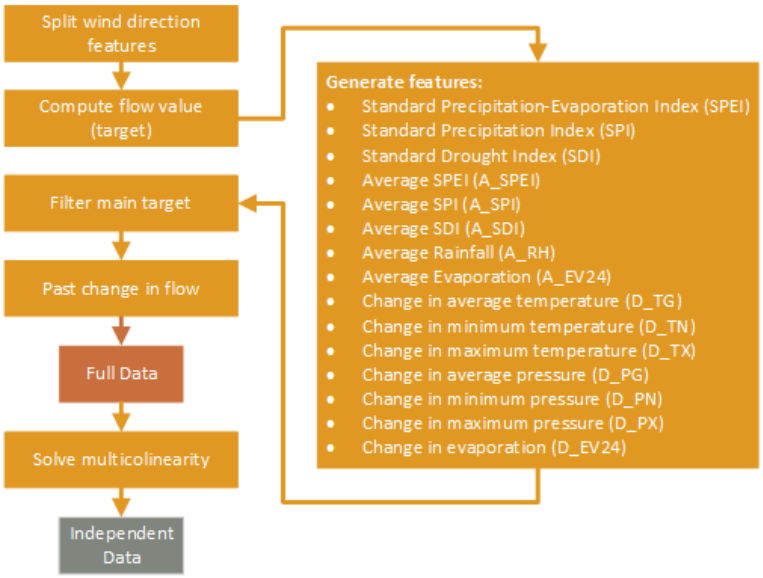

Figure 13: Feature Enigneering Steps

The first step is to split the wind direction features. Wind direction is expressed as the angle representing the direction from which the wind is blowing (Figure 19). New features will be generated that will be used to represent the North-South and East-West components of the wind direction using the sine and cosine functions respectively. Next the flow of the river is computed from the river height value using the formula provided by the WB and the WCG (Aa's and Provincie Drenthe 2019).

The next step is to generate new features using the rainfall and evaporation signals and also attempt to encompass the dynamics of the catchment area by creating difference (discrete derivatives) features of various original meteorological variables. First, the instantaneous values of the Standard Precipitation-Evaporation Index (SPEI), Standard Precipitation Index (SPI) and Standard Drought Index (SDI) are computed. Next the average of these features together with the averages of the rainfall and evaporation features is computed over a period of 1, 6 and 12 days. The next features to be generated are the changes in temperature, pressure and evaporation that have occured in the past 1, 6 and 12 days. The signal abreviation used for these features is explained in Table 1

Table 1: New feature descriptions

Feature Name	Feature Description
TI_<location>	Temperature interval at <location>
PI_<location>	Pressure interval at <location>
UI_<location>	Humidity interval at <location>
SPI_<location>	Standard Precipitation Index at <location>
SPEI_<location>	Standard Precipitation-Evaporation Index at <location>
A_<signal>_<location>_<time>	Average <signal> value at <location> in the past <time> days
D_<signal>_<location>(T+<time>)	Change in <signal> at <location> from <time> days ago

After this the flow values is filtered using a centered simple moving average to remove any noise in the final target. Also since the emphasis is placed on low flow samples, the distribution of the flow is split into 3 regions (low, medium and high flow). For each of these interval a different filtering windows is used with the lowest interval having the smallest window as to preserve the more information from the original signal. Finally, past changes in the flow value of the river are computed and the full data, containing all features to be used is obtained.

Before the FE step is completed, any multicolinearity in the resulting data set has to be resolved. Having multiple correlated features in the dataset does not reduce the predictive power of the model, but it will reduce the ease of interpretation of the final model. Having multiple similar feature to choose from, the importance of said features will be diminished. To solve this the variable inflation factor (VIF) is used, and features with the highest VIF score are sequentially removed until the highest VIF in the data set is less than 5, the equivalent of $0.8R^2$ score.

5 Modeling and Evaluation

In this section the steps taken in obtaining the final predictive models together with the most optimal features are explained

5.1 Feature Selection

Feature Selection (FS) is the process of selecting the most relevant subset when constructing the model. Selecting a smaller, more relevant number of features allows for a more simple, easy to interpret model, improved computational efficiency (shorter training times) and reduced overfitting (James, 2013; Bermingham, 2015). Two methods are taken in obtaining the optimal feature subset. First, Forward Feature Selection (FFS) is used to obtain the top 10 features, based on R^2 score, that results in the the best model performance (Figure 14). In FFS, features are sequentially added to the set of predictors according to the performance increase they provide, meaning that the feature that results in the best model performance after it has been included is selected.

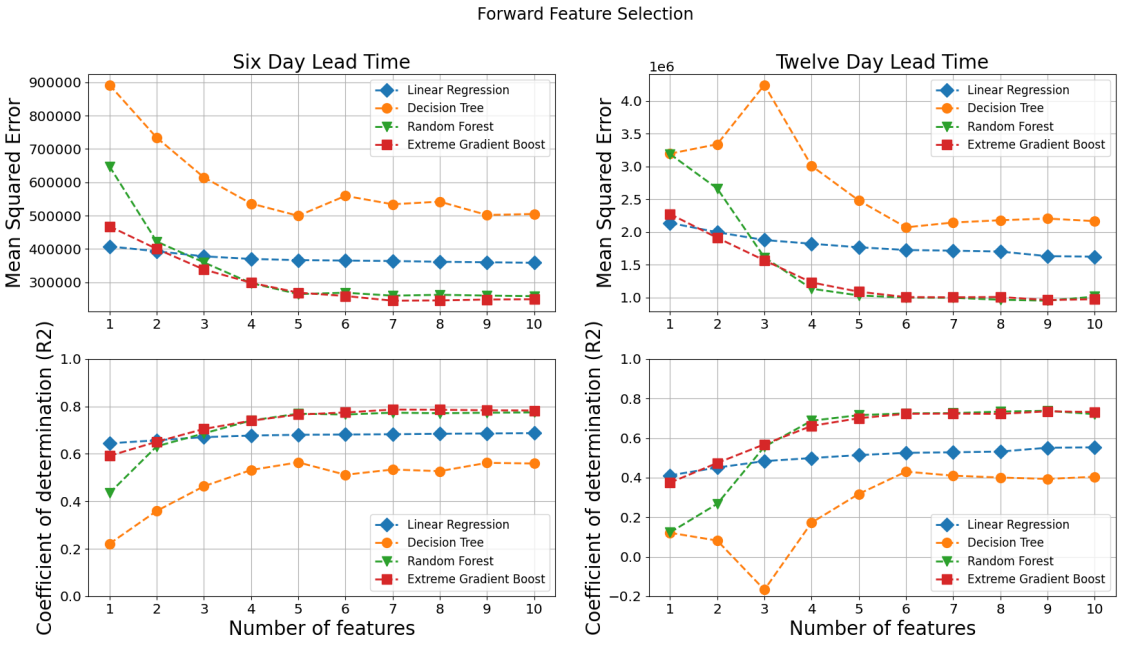

Figure 14: Forward feature selection results for all four models

After this step is complete, Exhaustive Feature Selection (EFS) is performed on the top features to obtain the best possible subset of features. In EFS, every possible feature combination from the resulting top 10 features is evaluated and the one with the smallest MSE is selected as the best subset. The results of the EFS is shown in Table 2.

Table 2: Exhaustive Feature Selection

6 Day Prediction				12 Day Prediction			
Linear Regression	**Decision Tree**	**Random Forest**	**Extreme Gradient Boost**	**Linear Regression**	**Decision Tree**	**Random Forest**	**Extreme Gradient Boost**
DR_280	MONTH	MONTH	MONTH	MONTH	MONTH	MONTH	MONTH
MONTH	A_SPEI_280_12	D_TN_280(T+6)	A_SPI_279_12	SPEI_279	D_TN_280(T+12)	A_SPEI_280_12	D_TN_280(T+12)
A_SPI_279_12	D_TN_280(T+12)	D_TN_280(T+12)	D_TN_280(T+12)	A_SPEI_280_12	D_TX_279(T+12)	D_TN_280(T+12)	D_TX_279(T+12)
D_TN_280(T+6)	D_TX_279(T+12)	D_TX_279(T+12)	D_TX_279(T+12)	A_SPI_279_12	D_EV24_280(T+12)	D_TX_279(T+12)	D_PX_280(T+12)
D_TX_279(T+12)	D_PX_280(T+6)	D_EV24_280(T+12)	D_PX_280(T+12)	D_TX_279(T+12)	DQ_S(T+1)	D_PX_280(T+12)	D_EV24_280(T+12)
D_PX_279(T+1)	D_EV24_280(T+12)	DQ_S(T+1)	D_EV24_280(T+12)	D_PX_280(T+6)	DQ_S(T+12)	D_EV24_280(T+12)	DQ_S(T+1)
D_PX_280(T+6)	DQ_S(T+1)	DQ_S(T+12)	DQ_S(T+1)	D_PX_280(T+12)	H_S	DQ_S(T+1)	DQ_S(T+12)
D_EV24_279(T+6)	H_S	H_S	DQ_S(T+12)	DQ_S(T+1)		DQ_S(T+12)	H_S
DQ_S(T+1)				DQ_S(T+12)		H_S	
DQ_S(T+12)				H_S			

5.2 Hyperparameter Tunning

Hyperparameters are model specific parameters that are used to control the learning (fitting) process as well as model complexity. More complex models provide better prediction performance, however they are prone to overfitting, simply memorizing the data instead of the relationships. Since DTs are the base estimators for RF and XGB, it is the mainly their parameters that will be tuned. The *maximum depth* of each tree represents the number of levels the tree will have. Whenever a split is made the depth increased by one, as such the greater the value of this parameter, the more complex the model will be. The *minimum samples per split* represents the minimum number of samples present of a particular node required to make a split, should a node have less samples present in it, then it will not perform a split. Complexity decreases as this parameter is increased. The *maximum number of lead nodes* represent the amount of end nodes (leafs) that the model will have, complexity increases as this parameter is increased. For the RF model, the same parameters as DT will be tunned, while the *number of estimators* parameters will be set arbitrarily high. For XGB besides the *maximum depth* of each tree, the *number of estimators* and *learning rate* will be tuned. The number of estimators controls how many DTs the models will use and increasing this parameters increases complexity. Learning rate represents how strong the model will successively correct error made in previous attempts, complexity increases proportionally with this parameter

Optimizing these parameters is the task of obtaining the optimal values for these parameters. To do this, each parameter together with a value range for that parameter is used to train and evaluate the model. For this step of the modeling stage, the models that provides the smallest mean squared error are chosen as optimal. The parameters chosen for the tuning step are presented in Table 3.

Table 3: Parameter tuning ranges

Model	Parameter Range
Decision Tree	max_depth: [1, 3, 5, 7, 9, 11, 13, 15, 17, 20] max_leaf_nodes: [2, 35, 68, 101, 134, 167, 200, 233, 266, 300] min_samples_split: [2, 35, 68, 101, 134, 167, 200, 233, 266, 300]
Random Forest	max_depth: [1, 2, 4, 6, 8] max_leaf_nodes: [2, 35, 68, 101, 134, 167, 200, 233, 266, 300] min_samples_split: [2, 57, 112, 168, 223, 278, 334, 389, 444, 500]
Extreme Gradient Boost	n_estimators: [2, 24, 46, 68, 90, 112, 134, 156, 178, 200] max_depth: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10] learning_rate: [0.005, 0.01, 0.015, 0.02, 0.025, 0.03, 0.035, 0.04, 0.045, 0.05]

5.3 Resulting Models

The results of the previous feature selection and tuning steps are presented in Table 4, the model with the best performance is RF.

Table 4: Model Performance

		Performance Measure	Linear Regression	Decision Tree	Random Forest	Extreme Gradient Boost
Flow Change	6 Day	R2	0.28	0.55	**0.65**	0.62
		MSE	843805	537274	**417002**	458768
		MAE	624	487	**425**	452
	12 Day	R2	0.12	0.26	**0.42**	0.37
		MSE	3330765	2835074	**2229604**	2435251
		MAE	1234	1130	**1000**	1053
Flow	6 Day	R2	0.95	0.98	0.98	**0.99**
		MSE	686394	437658	282486	**205708**
		MAE	546	429	342	**283**
	12 Day	R2	0.83	0.90	0.92	**0.95**
		MSE	2598790	1459592	1239533	**746442**
		MAE	1061	785	735	**533**

The RF outperforms all the other models on the change in flow prediction, however XGB provides better performance on the resulting flow. This is down to the fact that RF provides better performance for samples with higher drought weight (Figure 10) while XGB has similar error rate for low and high weight samples. The improved performance of RF compared to XGB could be down to the internal structure of the two models. RF relies on paralel ensembling of DT, and computes the final prediction values by averaging the predictions of all DT in it's ensemble, while XGB successively corrects the error of the current DT using another DT. Aggregation (averaging multiple "weaker" learners) has consistently proven to be a reliable method for deacreasing bias, as multiple learners are trained on a slightly different dataset, it manages to capture the essential relationship present as well as eliminating noise resulted from any of the weaker learners overfitness. Such a capability is not present in the XGB model, should the learner make a significant error at any particular step, that error will propagate throughout the entire model, thus resulting in an overall greater error for the entire distribution.

The resulting fit of the RF model can be seen in Figure 15. The two resulting models

have the following parameters:

$$RF_6 : min\_samples\_split = 2, max\_leaf\_nodes = 300, max\_depth = 6$$
$$RF_{12} : min\_samples\_split = 57, max\_leaf\_nodes = 200, max\_depth = 8$$

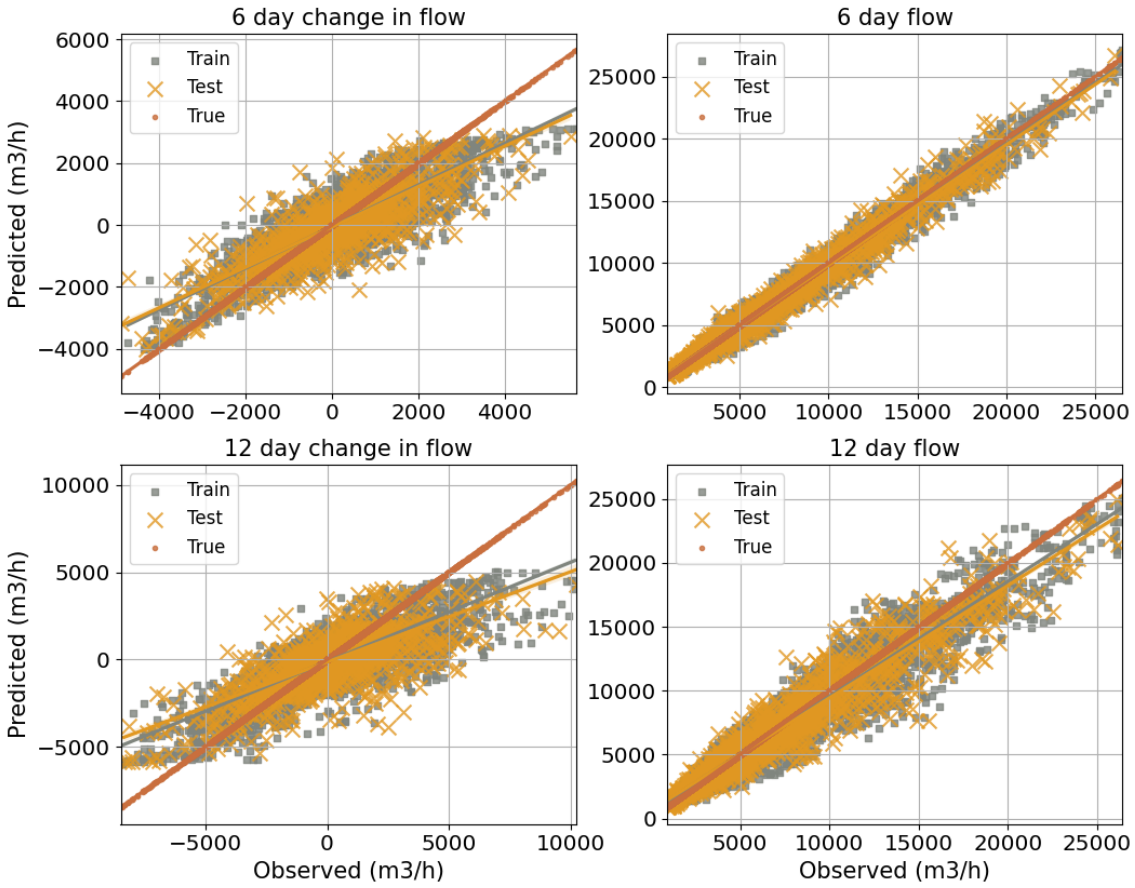

Figure 15: Random Forest fit for the 6 day prediction (left) and 12 day prediction (right)

6 Deployment and Feedback

In this section, the Shapley interpretation values will be presented for the two winning models, with six and twelve day ahead predictions. The Shapley dependence plots are presented in the annex.

6.1 Model Interpretations

Six day lead time

Figure 16 depicts the global interpretations for the six day model and local interpretation for periods of drought (low flow and/or fast decreasing flow). The most predominant feature for this model is the past change change in flow from one day ago (Figure 16f). Such dependence could reflect an slow inertial element of the river dynamic, meaning that the flow will not fluctuate significantly from one day to another.

Features that have to do with the heat present in the catchment area such as change in evaporation and maximum temperature (Figure 16e and 16d respectively) have a negative contribution to the flow evolution. In this aspect, it makes sense that as evaporation levels and maximum temperature increase, more water will evaporate from the river and therefore it is expected that the flow will decrease in such conditions. The negative contribution of the river height (Figure 16h) could represent a balancing mechanism of the river capacity. It makes sense that the less water in the river the lower the chances of a drastic river decrease, as there is no more water for the river to lose. Also as the river reaches its maximum capacity there is little possibility for the amount of water in the river to increase any further.

The positive effect of sudden increases in minimum temperature (Figure 16b) could represent a temporal characteristic, since it is expected that minimum temperature rise during the summer, when the flow is seen to increase drastically as rainfall amounts start to increase (Figure 4). The current month in which the prediction is made (Figure 16a) indicates that over a period of 6 days, the flow is expected to decrease in the months of April, May and June, while for the months of September, October, November and December the flow is expected to increase. For the remaining months the flow doesn't show a any particular increasing or decreasing behaviour.

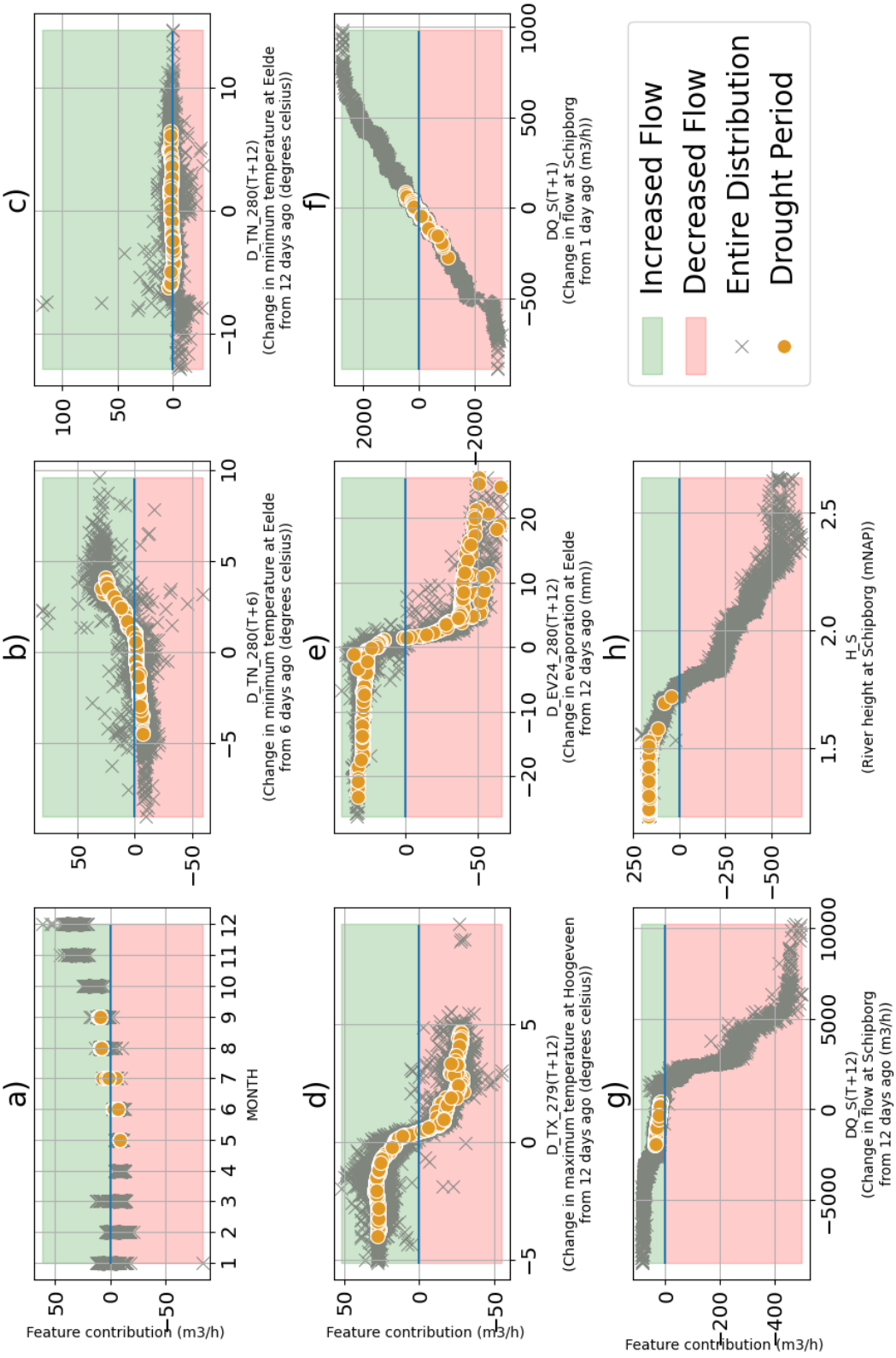

Figure 16: Six day prediction model global and drought period interpretations

Twelve day lead time

Figure 17 depicts the global interpretations for the twelve day model and local interpretation for periods of drought (low flow and/or fast decreasing flow). Similar to the six day model, the effect of change in evaporation (Figure 17f) and maximum temperature (Figure 17d) have a negative relation with the change in flow, however at a higher magnitude. Unlike for the 6 day model, the effect of minimum temperature (Figure 17c) seems to be reversed, with sudden decreases indicating an increase in flow. Another similarity with the 6 day model is the predominant effect of the previous day change in flow (Figure 17g) as well as the change in flow from 12 days in the past (Figure 17h), with the former having similar magnitude and the later have a greater magnitude.

A peculiar effect can be seen in the Precipitation-Evaporation index (Figure 17b), with wet periods indicating a slight decrease in flow. This could be a consequence of the choice in averaging interval (12 days for SPEI), since the maximum travel time for the water from the most distant parts of the river is 3 days, this feature could be capturing the ending of rainfall events, as no more water is added into the system and the river begins to stabilize.

Changes in pressure (Figure 17e) seem to have an increasing effect only when these changes are sudden, with more stable pressures having little effect on the flow. The month in which the prediction is made (Figure 17a) has a similar effect to the first model, with spring and summer months (March, April, May, June and July) showing a negative effect on the flow, while for the rest of the year, the flow is expected to increase with the winter period showing the most drastic changes.

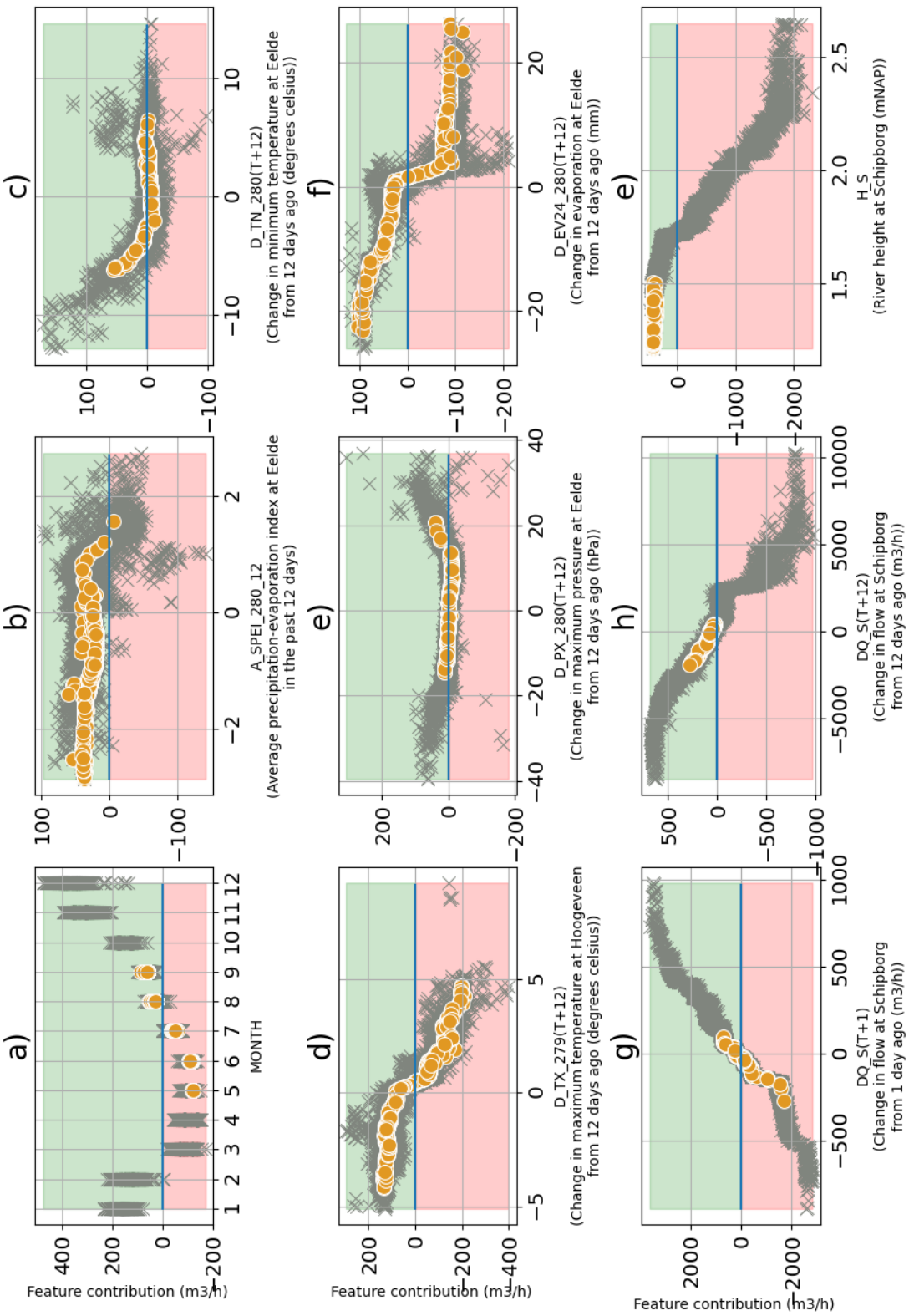

Figure 17: Twelve day prediction model global and drought period interpretations

6.2 Discussion

When it comes to the performance of both models, better results can be achieved using a time series approach as proven by the overall higher Shapley values of both models for past changes in flow. This comes at a cost for interpretability, given that besides an inertial behaviour (if it decreased today, it will decrease in six/twelve days) there is no other possible explanation for the effect of these features. It could be possible to achieve better performance by adopting a full timeseries approach and ditch any other meteorological features completely, however this comes a the cost of complete loss of interpretability.

For samples depicting periods of drought (orange dots in Figures 16 and 17) the change in flow is affected by features describing the heating of the catchment area. Maximum temperature could have an indirect effect through rainfall, as it is typical for the region to experience a drop in temperatures as there is more rain on a daily basis. Evaporation is the most direct effect on the river out of all of the features used, as it explains the amount of water lost by the surrounding environment during that day. The lack of features representing rain is interesting as one would expect the effect of rain to be more prominent. However, since the effect of any rainfal will be observed with a certain delay, mainly because of the fact that rain from near Hoogeveen has a transitory period of about 3 days and also the time it takes water to seep into the river from the surrounding ground. Since no lagging has been applied to rain signals it makes sense that there is no instantaneous effect detected in the analysis.

The most surprising feature is the average precipitation-evaporation, as it seems to have a constant effect until values that depict wetter periods, when the effect becomes negative. This is counter-intuitive, as it is expected that wetter periods translate into more water present in the system, but according to the twelve day model, wetter periods result in a slight drop in the amount of water.

Minimum temperature for the twelve day model indicates that sudden decreases in temperatures result in increasing flow. This feature could take effect during the final days of the drought period as temperatures begin to increase for the upcoming autumn.

7 Artifact

The artifact of this project will be a software application that will allow the user to Update the current dataset, train the model on the newly gathered dataset, make a prediction using the model, and provide the Shapley values for the training dataset as well as the predictions made.

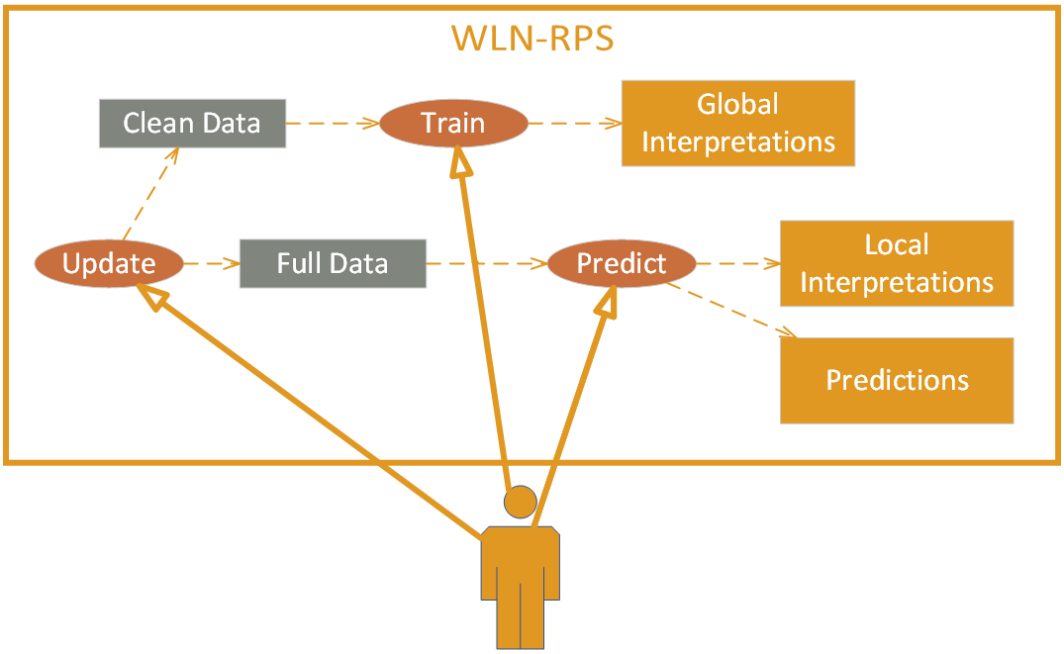

Figure 18: Use case diagram of the WLN-RPS application

Figure 18 represent the use case diagram of the application. It will have three functions, update, train and predict.

The updating function will fetch, process and generate features required to train the model, with the output of this function being the clean data set (data used for training) and the full dataset (data used for predicting). The difference between these two datasets is that because of the filtering performed in the processing step, the final sample of the clean dataset will be missing, while these values are kept in the full dataset.

The training function uses the feature resulted from the feature selection step together with the tuned parameters to obtain the models required for making the prediction, together with the shapley values of the enitre dataset (with the known outcome).

The prediction function will use the last 6 day values of the selected feature for both models to obtain the required predictions as well as provide the shapley values that will provide the interpretation for the resulting predicted value.

Global interpretation will help in understanding how the model makes predictions, while the hope is that the local interpretations will provide the justification required in case the decision to start downscaling is made. Information from global interpretations will help understand and validate the local predictions that the models makes.

The application will have a simple user interface that will provide the user with a button for each function, as well as an input field for the current capacity of the plant to display the plant downscaling possibility according to the imposed procedure (Figure 8). The output of each function will be a csv table that will be used by the company to visualize the interpretation and river evolution using Tableau.

8 Conclusion

This paper aimed at predicting the streamflow values for the Drentsche Aa river six and twelve days in advance and provide interpretation for the model predictions. Using expert knowledge new features were generated that would help capture more complex relationships present in the data. The most relevant features according to model performance were identified using feature selection methods. Four models were tested, Linear Regression, Decision Tree, Random Forest and Extreme Gradient Boost, with Random Forest providing the best performance for the change in flow. Interpretations were provided using the Shapley values method that explains the contribution of each feature to the final prediction values.

Should a solution like this be implemented in a real world application, it is important to establish early on in the development of the product how much emphasis will be placed on the model. Should the model be seen as an "expert" in the river dynamics, then emphasis should be placed on it's predictive performance in order for it to have more accuracy on the predictions. This would give the model a more direct role in the final decision to begin downscaling, as the predicted value that it gives will have to be taken for granted. To this extend, and also because of the fact that the river has a rather slow evolution, better precision can be achieved by implementing a solution close to a time series approach. This can be achieved by introducing more features describing the past evolution of the river, however would lead to a decrease in interpretability because of the difficulty in explaining the effect of these features. On the other hand, if the model should have a more supportive role, that of explaining why it thinks the river will behave the way it does, then focus should be placed on it's interpretation capabilities. River evolution could be better explained and interpreted by completely dropping any past flow features in an attempt to predict the streamflow using only catchment characteristics. Such a solution will serve the decision makers with empirical reasoning on what causes changes in the current flow. That way the decision to begin downscaling will be in the hand of the decision makers, with the model serving as a way to give a more empirical reasoning behind the choice.

References

World Health Organization (2017). *Guidelines for drinking-water quality*. OCLC: 975491910.

Agency, Environmental Protection (n.d.). "Factoids: drinking water and ground water statistics for 2007. March 2008, April 2008". In: ().

Yevjevich, Vujica M., L. Veiga da Cunha, and Evan Vlachos, eds. (2012). *Coping with droughts*. Classic resource edition. Highlands Ranch, Colorado, USA: Water Resources Publications, LLC. 417 pp.

Rossi, G., L. Castiglione, and B. Bonaccorso (2007). "Guidelines for Planning and Implementing Drought Mitigation Measures". In: *Methods and Tools for Drought Analysis and Management*. Ed. by Giuseppe Rossi, Teodoro Vega, and Brunella Bonaccorso. Vol. 62. Series Title: Water Science and Technology Library. Dordrecht: Springer Netherlands, pp. 325–347.

Bazza, Mohamed (2002). "Water Resources Planning and Management for Drought Mitigation". In:

Cancelliere, A. et al. (Apr. 23, 2007). "Drought forecasting using the Standardized Precipitation Index". In: *Water Resources Management* 21.5, pp. 801–819.

Rossi, Giuseppe, Teodoro Vega, and Brunella Bonaccorso, eds. (2007). *Methods and tools for drought analysis and management*. Water science and technology library 62. OCLC: 255693824. Dordrecht: Springer. 418 pp.

FAO (2001). "Inferences of a Drought Mitigation Action Plan. Expert Consultation and Workshop on Drought Preparedness and Mitigation in the Near East and the Mediterranean, Organized by FAO/RNE, ICARDA and EU". In:

Dziegielewski, B. (2003). "Long-Term and Short-Term Measures for Coping with Drought". In: *Tools for Drought Mitigation in Mediterranean Regions*. Ed. by Giuseppe Rossi et al. Red. by V. P. Singh. Vol. 44. Series Title: Water Science and Technology Library. Dordrecht: Springer Netherlands, pp. 319–339.

Mitchell, Tom M. (1997). *Machine Learning*. McGraw-Hill series in computer science. New York: McGraw-Hill. 414 pp.

Koza, John R. et al. (1996). "Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming". In: *Artificial Intelligence in Design '96*. Ed. by John S. Gero and Fay Sudweeks. Dordrecht: Springer Netherlands, pp. 151–170.

Wang, Lili et al. (2019). "Improving the prediction accuracy of monthlystreamflow using a data-driven model based on a double-processing strategy. J.Hydrol." In:

Rezaie-Balf, Mohammad et al. (May 2019). "Daily river flow forecasting using ensemble empirical mode decomposition based heuristic regression models: Application on the perennial rivers in Iran and South Korea". In: *Journal of Hydrology* 572, pp. 470–485.

Gauch, Martin and Jimmy Lin (June 5, 2020). "A Data Scientist's Guide to Streamflow Prediction". In: *arXiv:2006.12975 [physics, stat]*. arXiv: 2006.12975.

Meng, Erhao et al. (Jan. 2019). "A robust method for non-stationary streamflow prediction based on improved EMD-SVM model". In: *Journal of Hydrology* 568, pp. 462–478.

Liu, Zhiyong et al. (Oct. 16, 2015). "A multivariate conditional model for streamflow prediction and spatial precipitation refinement". In: *Journal of Geophysical Research: Atmospheres* 120.19.

Wen, Xiaohu et al. (Mar. 2019). "Two-phase extreme learning machines integrated with the complete ensemble empirical mode decomposition with adaptive noise algorithm for multi-scale runoff prediction problems". In: *Journal of Hydrology* 570, pp. 167–184.

Zhang, Duo, Geir Lindholm, and Harsha Ratnaweera (Jan. 2018). "Use long short-term memory to enhance Internet of Things for combined sewer overflow monitoring". In: *Journal of Hydrology* 556, pp. 409–418.

Hussain, Dostdar and Aftab Ahmed Khan (Sept. 2020). "Machine learning techniques for monthly river flow forecasting of Hunza River, Pakistan". In: *Earth Science Informatics* 13.3, pp. 939–949.

Parisouj, Peiman, Hamid Mohebzadeh, and Taesam Lee (Oct. 2020). "Employing Machine Learning Algorithms for Streamflow Prediction: A Case Study of Four River Basins with Different Climatic Zones in the United States". In: *Water Resources Management* 34.13, pp. 4113–4131.

Tongal, Hakan and Martijn J. Booij (Sept. 2018). "Simulation and forecasting of streamflows using machine learning models coupled with base flow separation". In: *Journal of Hydrology* 564, pp. 266–282.

Asefa, Tirusew et al. (Mar. 2006). "Multi-time scale stream flow predictions: The support vector machines approach". In: *Journal of Hydrology* 318.1, pp. 7–16.

Tyralis, Hristos, Georgia Papacharalampous, and Andreas Langousis (Apr. 30, 2019). "A Brief Review of Random Forests for Water Scientists and Practitioners and Their Recent History in Water Resources". In: *Water* 11.5, p. 910.

Chen, Tianqi and Carlos Guestrin (June 10, 2016). "XGBoost: A Scalable Tree Boosting System". In: *arXiv:1603.02754 [cs]*. arXiv: 1603.02754.

Fan, Junliang et al. (May 2018). "Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China". In: *Energy Conversion and Management* 164, pp. 102–111.

Xiao, Qingyang et al. (Nov. 20, 2018). "An Ensemble Machine-Learning Model To Predict Historical PM $_{2.5}$ Concentrations in China from Satellite Data". In: *Environmental Science & Technology* 52.22, pp. 13260–13269.

Zhang, Rong et al. (May 2019). "Meteorological drought forecasting based on a statistical model with machine learning techniques in Shaanxi province, China". In: *Science of The Total Environment* 665, pp. 338–346.

Chen, Xing et al. (Jan. 5, 2018). "EGBMMDA: Extreme Gradient Boosting Machine for MiRNA-Disease Association prediction". In: *Cell Death & Disease* 9.1, p. 3.

Xia, Yufei et al. (July 2017). "A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring". In: *Expert Systems with Applications* 78, pp. 225–241.

Ni, Lingling et al. (July 2020). "Streamflow forecasting using extreme gradient boosting model coupled with Gaussian mixture model". In: *Journal of Hydrology* 586, p. 124901.

Hyndman, Rob J. and Anne B. Koehler (Oct. 2006). "Another look at measures of forecast accuracy". In: *International Journal of Forecasting* 22.4, pp. 679–688.

Shapley, Lloyd S. (1951). "Notes on the n-person game II: the value of an n-person game". In:

Roth, Alvin E., ed. (Oct. 28, 1988). *The Shapley Value: Essays in Honor of Lloyd S. Shapley*. 1st ed. Cambridge University Press.

Lundberg, Scott M., Gabriel G. Erion, and Su-In Lee (Mar. 6, 2019). "Consistent Individualized Feature Attribution for Tree Ensembles". In: *arXiv:1802.03888 [cs, stat]*. arXiv: 1802.03888.

Aa's, Waterschap Hunze en and Waterbedrijf Groningen en Provincie Drenthe (2019). *TOPSOIL DUURZAME WATERKWALITEIT DRENTHE*.

ANNEX

Name	Description
YYYYMMDD	Date (YYYY=year MM=month DD=day)
DDVEC	Vector mean wind direction in degrees (360=north, 90=east, 180=south, 270=west,
FHVEC	Vector mean windspeed (in 0.1 m/s)
FG	Daily mean windspeed (in 0.1 m/s)
FHX	Maximum hourly mean windspeed (in 0.1 m/s)
FHXH	Hourly division in which FHX was measured
FHN	Minimum hourly mean windspeed (in 0.1 m/s)
FHNH	Hourly division in which FHN was measured
FXX	Maximum wind gust (in 0.1 m/s)
FXXH	Hourly division in which FXX was measured
TG	Daily mean temperature in (0.1 degrees Celsius)
TN	Minimum temperature (in 0.1 degrees Celsius)
TNH	Hourly division in which TN was measured
TX	Maximum temperature (in 0.1 degrees Celsius)
TXH	Hourly division in which TX was measured
T10N	Minimum temperature at 10 cm above surface (in 0.1 degrees Celsius)
T10NH	6-hourly division in which T10N was measured; 6=0-6 UT, 18=12-18 UT, 24=18-24 UT
SQ	Sunshine duration (in 0.1 hour) calculated from global radiation (-1 for <0.05 hour)
SP	Percentage of maximum potential sunshine duration
Q	Global radiation (in J/cm2)
DR	Precipitation duration (in 0.1 hour)
RH	Daily precipitation amount (in 0.1 mm) (-1 for <0.05 mm)
RHX	Maximum hourly precipitation amount (in 0.1 mm) (-1 for <0.05 mm)
RHXH	Hourly division in which RHX was measured
PG	Daily mean sea level pressure (in 0.1 hPa) calculated from 24 hourly values
PX	Maximum hourly sea level pressure (in 0.1 hPa)
PXH	Hourly division in which PX was measured
PN	Minimum hourly sea level pressure (in 0.1 hPa)
PNH	Hourly division in which PN was measured
VVN	Minimum visibility; 0: <100 m, 2:200-300 m,..., 49:4900-5000 m, 50:5-6 km, 56:6-7 km, 57:7-8 km,..., 79:29-30 km, 80:30-35 km, 81:35-40 km,..., 89: >70 km)
VVNH	Hourly division in which VVN was measured
VVX	Maximum visibility; 0: <100 m, 2:200-300 m,..., 49:4900-5000 m, 50:5-6 km, 56:6-7 km, 57:7-8 km,..., 79:29-30 km, 80:30-35 km, 81:35-40 km,..., 89: >70 km)
VVXH	Hourly division in which VVX was measured
NG	Mean daily cloud cover (in octants, 9=sky invisible)
UG	Daily mean relative atmospheric humidity (in percents)
UX	Maximum relative atmospheric humidity (in percents)
UXH	Hourly division in which UX was measured
UN	Minimum relative atmospheric humidity (in percents)
UNH	Hourly division in which UN was measured
EV24	Potential evapotranspiration (Makkink) (in 0.1 mm)

Figure 19: KNMI Signal Descriptions