A Non-Parametric Bayesian Network Hydrologic Model

A Case Study of a Lowland Catchment

Sjoerd Gnodde

Witteveen Bos

TU Delft

A Non-Parametric Bayesian Network Hydrologic Model

A Case Study of a Lowland Catchment

A Non-Parametric Bayesian Network Hydrologic Model

A Case Study of a Lowland Catchment

MSc Thesis

to obtain the degree of Master of Science
at Delft University of Technology,
to be defended online on June 29, 2020 at 14:00[1].

by

Sjoerd Gnodde

[1] Please contact the author if you want to attend the defence.

Thesis committee:

dr.ir. O. Morales-Nápoles,	TU Delft (Chairman)
dr. M. Hrachowitz,	TU Delft
dr. E. Ragno,	TU Delft
ir. B. Dekens,	Witteveen+Bos
dr.ir. J. Hoch,	Universiteit Utrecht

Keywords:	Lowland catchments, Hydrologic modeling, Non-parametric Bayesian Networks, Data-driven models, Model optimization
Cover design	Anniek Keijer & Sjoerd Gnodde
Photo back cover	The river Vledder Aa during the Water-Op-Maat project middenloop Vledder Aa. Photo by Corné Joziasse.

An electronic version of this thesis is available at
http://repository.tudelft.nl/.

I believe that we do not know anything for certain,
but everything probably.

Christiaan Huygens

Contents

SUMMARY

For many years, improvements have been made in hydrologic modelling in catchments. For a long time, a vast amount of data has been collected, which contributes to the quality of these hydrologic models. Since the last decades, this has been complemented by various satellite measurements. This makes way for a new, data-driven generation of models. This research proposes a non-parametric Bayesian network (NPBN) to model hydrologic processes. The Bayesian network (BN) is a directed, acyclic graph (DAG), in which the variables are represented by the nodes and the conditional probability distribution between variable pairs is represented by the arcs. This graphically models a complex implementing of Bayes' theorem. BNs are widely used in complex risk processes, due to the ability to handle complex probabilistic dependencies. More specifically, NPBNs are computationally less expensive than many conceptual hydrologic models and are flexible to handle different continuous data sources. Moreover, variables not directly related to either a water flux or quantity, can be implemented directly into an NPBN. An unsaturated BN with an effective layout, which is the model that is used in this thesis, can predict other variables than the target as well, keeps probability distributions although other variables are fixed and models conform physical relations between variables. Other data-driven methods, among which is the saturated BN, do not have these advantages. The goal of this thesis is to make an NPBN for a lowland catchment and testing its performance. This boils down to the following research question:

> What is the optimal setup of a Bayesian network hydrologic model in a lowland catchment, and how does it perform?

The selected case study is the catchment of the Vledder, Wapserveense and Steenwijker Aa, which is located in the north of The Netherlands. This catchment makes this research the first one in which an NPBN is comprehensively implemented for (1.) a single catchment in which the catchment processes are modelled, (2.) in a Dutch catchment and (3.) a lowland, partially managed, catchment.

For the BN model, the following variables are selected and data is acquired: precipitation, temperature, solar radiation, soil moisture, NDVI, groundwater levels from a single well and surface water levels from a single measurement station. The input data combines terrestrial measurements with satellite measurements. Additionally, the monthly maximum daily average discharge (MMDAD) are selected as target variable. This target variable is used to optimise a number of parameters of the BN throughout this research, with the performance parameter Kling-Gupta efficiency (KGE). The quality of the data is looked into. The data could not be rejected on the basis of the Budyko framework. No further decisive tests on the data could be performed, except for some minor filtering of impossible and implausible values. However, using imperfect data often reflects real-world application of such a model.

In this thesis, the testing has been done predominantly a posteriori. This is because the central goals is to make the model predict the MMDAD well and the other goals for the model are secondary. Therefore, a complete BN is introduced early in the research. This model is used to analyse all the parameters in the model, which is mostly done through calculation of the KGE, a performance indicator that is mostly used in hydrology, of the prediction of the MMDAD.

The method to use continuous variables in a BN that is used in this thesis, is based on copulas. These are multivariate probability space constructed with marginals constructed with the cumulative distribution functions (CDFs) of the variables. A wide range of copula types exist. For this thesis, the Gaussian copula was selected to be implemented for all variable combinations, because this type has no tail dependence and allows for the use of the multivariate normal distribution to calculate a conditioned copula. The correlations that are used, are the partial, normal rank correlations. This thesis tests the assumption that the Gaussian copula is sufficient for all variable combinations, in context to the performance of other copulas. The Gaussian copula is not optimal for all combinations, but does not make for a highly unsatisfying fit. Moreover, this method is far more convenient than using the alternative method called vine-copula method and most likely gives a better fit than the other alternative: the combined probability distributions than the usage of a single Archimedean copula. An exception to the latter is that the Frank copula gave better fits to several of the variable pairs than the Gaussian copula.

Three distributions are compared to model the marginal distributions. These are (1.) the empirical cumulative distribution function (ECDF), which is a step function, (2.) an altered logistic function, which is introduced in this thesis, and (3.) the Gaussian mixture model. The latter has been selected as this performed equally well as the ECDF when implemented for all variables in predicting the MMDAD and has the possibility to predict data points that are not in the data yet. This turned out to be only marginally the case. Therefore, a novel alteration function has been proposed to shift the predictions to make the model predict more extrapolated events. In the case of the model proposed in this thesis, a minor shift has been used to predict peak discharges better.

A threefold of parameters of the BN has been analysed and a favourable setting has been chosen: first, it was considered using the median or the mean of the predicted variable. Using the mean was considered slightly more favourable. Then, it was analysed how many samples to use for a prediction, of which 500 samples were selected. Finally, is was analysed how many days prior to the MMDAD event to take into account for aggregating the other variables. Here, 8 days were selected.

Additionally, a sensitivity analysis has been performed to understand what influences of artificial errors would be. Because usually many datapoints are aggregated per variable, random errors have a marginal influence on the network. Systematic errors do not have any influence on a BN, except for that a predicted variable is off by that systematic error. New systematic errors have a bigger influence. For these errors, the sign (plus or minus) of the correlation in combination with the sign of the new systematic error matters: when this is the same, the error is generally less.

Criteria for a practical, well-performing BN have been presented and a strategy to create such a model that satisfies these criteria has been assembled on the basis of the characteristics of a BN. This strategy left the selection of some connection implementations up for interpretation. Which implementation is chosen, has been decided by looking at which had the best prediction of a relevant variable within the network.

The final model gave a median, k-fold tested KGE of 0.73 when predicting the MMDAD.

The larger discharges are slightly underestimated, however, this effect is less due to the shift function. It is also analysed how well the MMDAD is predicted, if not all other variables are fixed. Other variables can also be predicted, but not as well as the MMDAD.

Another novelty is that a BN model is benchmarked against a SOBEK model, a neural network, and a multiple linear regression model. The SOBEK model that is in use for flood prediction, performs less than the BN. The Neural Network most likely needs more data than is provided in this dataset and therefore also does not perform as well as the BN. Using a multiple linear regression is very easy and quick and performs just a bit less than the BN. However, all these other models lack some advantages that the unsaturated BN has.

Preface

Finishing of a Masters is like the queen stage of a grand tour. You've finished all of the other stages during the Bachelor and the beginning of the Master. Some were more difficult than others, but one thing is certain: the biggest hurdle is the final one. At the same time, this is also the stage in which you can show everything you've learned so far. And it's most definitely a long stage. Therefore, I took a significant time to determine my thesis subject. I wanted it to fit my interests perfectly: combining hydrologic models with data science and statistical models, and a little challenging. This would ensure such that I wouldn't get saddle pain halfway and make sure I can show what I'm capable of. I also decided that I wanted to join up with Witteveen+Bos, such that I could work in an inspiring setting, could ask colleagues for help, could work an a case study in which I was in direct contact with the managers of the area and in which the model I built could actually be implemented.

Now it was time to start the stage. During this stage it is always important to stay in front, work hard, but not too hard as this might come back to haunt you later. Despite this, some intermediate sprint were done in the run-up to the progress meetings. Although you finish alone, even making a thesis is a team sport. I would like to thank this team very much, because without you my thesis would have been a far less successful individual time trial. My core teammates existed of the thesis committee. First of all, I would like to thank Elisa Ragno for our regular meetings that made sure that I was getting the hang of this difficult subject. I want to thank Oswaldo Morales-Nápoles for chairing the committee despite being from a different department and taking time to help me with the technical part when Elisa wasn't available. I also want to thank Jannis Hoch for staying on my committee whilst switching jobs and making sure my academic writing was up to par. Moreover, I also want to Markus Hrachowitz for his advice on how to make a statistical model that complies with the physical reality. Then, I also want to thank Bart Dekens for always being available for smaller or larger questions and having been a pleasant colleague. Furthermore, I also want to thank all of my other colleagues at Witteveen+Bos. In particular, I would like to thank Bart van Es for helping me with the method. From Waterschap Drents Overijsselse Delta, I would like to thank Jelle de Jong and Zwannie Visser for letting me use their data and always being available for questions about the catchment and the data. Lastly, I really want to thank Rogier Dinkla for helping me with the writing, and Anniek Keijer for some magnificent designs.

During long stages, there is always the chance of a mechanical problem. In my case, this was the unforeseen coronacrisis. A quick bike change meant that I had to work from different places than the office or home for a long time, but eventually I got the hang of this as well. The final, difficult climb of the thesis was done at home, where I'm currently in the final sprint and have just passed the flamme rouge. Looking over my shoulder, I now have a great view of what's been my thesis, and further back, also my student years. What a ride it has been.

Sjoerd Gnodde
The Hague, June 2019

1

Introduction

1.1. Problem statement and objective

Finding the probability of future events has always been one of humanity's greatest efforts. Especially for areas close to rivers, identifying the chance of a flood can be vital. Therefore, conceptual hydrologic and rainfall-runoff models have been introduced to predict discharges in rivers from precipitation and other variables in the catchment area. Nowadays, as more and more people live in deltas and in the close proximity of rivers, these are increasingly important. Furthermore, as a consequence of climate change the frequency and intensity of heavy precipitation events has likely increased in Europe and North America (IPCC, 2014). These circumstances and developments increase the hazards of flooding.

The Netherlands is a country with many small and several large rivers. Because of its flat landscape, low elevation, high economic activity and high population, industry and agriculture density, consequences of flooding are big in the Netherlands. This means that reliable, flexible hydrologic models are of great importance to prevent chances of flooding.

Conventional models in hydrology, such as the HBV model (Bergström, 1976) are based on time-dependent storages and fluxes. Another range of models, often used for (mostly partially or fully managed) systems in the Netherlands, combine rainfall-runoff input with the Saint-Venant equations for modelling the hydrology in an area. For complex systems, both methods can become computationally heavy, because sufficiently small timesteps are needed to make the methods numerically stable. Moreover, they can also become difficult to implement for the user, because he or she has to make assumptions about atmospheric, lithospheric, biospheric and hydrospheric parameters.

Currently, every year more data is collected in the world than ever before (O'Dea, 2020), which also holds true for meteorological and hydrological data. This paves the way for novel, statistical, data-driven models in hydrology.

This research proposes non-parametric Bayesian networks (NPBNs) to model a hydrologic process. Bayesian networks (BNs) are widely used in complex risk processes, due to the ability to handle complex probabilistic dependencies. More specifically, NPBNs are computationally less expensive than conceptual models and are flexible to handle different continuous data sources. In conceptual models, for example, only water fluxes and storage can be part of the model. Other variables have to be translated to either a flux or a storage. In a Bayesian network, variables not directly related to either a water flux or quantity, can be implemented directly. For example, solar radiation can be used in a Bayesian network, whereas a conceptual model requires translation and combination to a flux, such as evapotranspiration. On top of that, BNs require only a simple, straightforward combination of variables with logical relations to construct a model. The rest of the probabilistic relations is fitted without the need to quantify more information than that is given with a layout and the order of dependence. In contrast to conceptual models, a number of variables that are difficult to measure, such as effective soil porosity and friction parameters over the whole catchment area, do not have to be measured or induced from a fit to data. Lastly, BNs use probability distributions throughout the network. By inference of observations, the masses of the other probability distributions are reordered. This means that it is possible to get a result from the model when not all variables have a known value, and that the final result (of all variables) is not just a number, but a whole probability distribution. This makes, for example, an uncertainty analysis possible. Therefore, Bayesian networks and related methods have slowly been introduced in hydrology over the previous two decades (e.g. Couasnon, 2017; Favre et al., 2004; Molina et al., 2005; Nasr et al., 2018; Paprotny and Morales-Nápoles, 2017; Sanjaya, 2018; Torres Alves, 2018; Yang H. et al., 2002).

Another data-driven method is, for example, the (artificial) neural network (NN). This is a powerful, novel method that is used in e.g. autonomous vehicles and in big data recommendation systems. However, for use as hydrologic models, NNs have some key disadvantages in comparison to the Bayesian network. With large NNs it is difficult to comprehend the inner workings for users. Unexpected results cannot be clarified and calculations in its hidden layers do not always make sense physically. In BNs on the other hand, the influence of each variable on the others can be instantly and easily interpreted with its (partial) correlation coefficients. The BN works related to (and therefore gives some insight in) the actual processes in the catchment. Most notably, for governmental bodies, models that are difficult to comprehend, are highly unfavourable, as they usually have to clarify their decision making process to their inhabitants.

Despite their great potential, BNs have some drawbacks as well. To begin with, BNs often have a less accurate prediction of a single value (mean or median of a distribution) than which can generally be expected from common hydrologic models, although specific research into this subject has not yet been conducted. This research does make a comparison between a BN and a conventional model, which in this case a is SOBEK model (see Section 9.2.2). Secondly, BNs can potentially generate very inaccurate results in a number of cases. The reason of this is not always certain, such as in Sanjaya (2018) and Torres Alves (2018), which makes the usage of BNs a greater risk. Moreover, NPBNs usually require a significant amount of data for their quantification. Because of the fact that relations are created solely from the data in these types of BNs, an extensive dataset that includes the same timesteps for all variables, has to be available. Lastly, BNs are not time dependent. Therefore, storage in the system has to be explicitly implemented in the network, if this has a significant impact on the catchment. As it is nearly impossible to measure the complete storage, or derive runoff from current events, Bayesian networks usually work with a lower frequency with aggregated values, such that many storage differences can be neglected.

The model proposed in this research, deals with all these shortcomings. The target variable in this research is the *monthly maximum daily average discharge (MMDAD)* out of a catchment. In the case of the main case study, this catchment is the Vledder, Wapserveense and Steenwijker Aa (see Chapter 3). The selection of this target variable is an optimum between having enough datapoints, sufficient data aggregation, and being an interesting variable. High discharges often result in hindrance of flooding downstream. For more argumentation about this variable, see Section 3.3.2.

The MMDAD is quantitative data, which calls for a quantitative model. Therefore, a so-called *non-parametric Bayesian network* is used. This is a model in which probability distributions of continuous variables are connected via probabilistic units called *copulas*.

Concluding, the goal of this research is to create an optimal Bayesian network for a lowland catchment in the Netherlands, test its performance, and benchmark its performance against a SOBEK model. This boils down to the following general research question:

> What is the optimal setup of a Bayesian network hydrologic model in a lowland catchment, and how does it perform?

1.2. Previous research into Bayesian networks and similar methods in hydrology

The term Bayesian network, and a description of its characteristics, was first introduced by Pearl (1985). This only contained discrete variables. Conditional probabilities between connected variables had to be explicitly defined.

Prior to this, copulas, have been introduced by Sklar (1959). Copulas can be used to link multiple empirical, continuous distributions and are used in this thesis as well. Over time, multiple types of copulas have been created, of which many are featured in Nelsen (2006), in an introductory manner.

NPBNs have first been introduced by Kurowicka and Cooke (2005), sampling vines has been described by Kurowicka and Cooke (2007) and the method has further been improved upon by Hanea et al. (2015). The main software package to use NPBNs is UniNet. This programme can handle NPBNs that work with discrete or continuous variables, or a combination of both. It was initially developed at the TU Delft by lead developer Dan Ababei and is now available from the website of his company LightTwist[1] with a free academic licence or a paid professional licence. Cooke et al. (2007) introduces and describes UniNet.

The method of conditioning a BN based on the Gaussian copula with a conditional multivariate normal distribution (see Section 2.6), has been introduced by Hanea et al. (2006). The conditional multivariate normal distribution itself has first been described by Eaton (1983).

Bayesian statistics were first introduced in hydrology by, among others, Krzysztofowicz (2002). This still had a typical hydrologic model in its core and mainly used Bayesian statistics for the probability distribution of the precipitation input and the river stage output. In the same year, Yang H. et al. (2002) has created a Bayesian network to model desertification. This is a model that also features societal and land use changes, and has discrete states for each of the

[1] `https://lighttwist-software.com/uninet/`

variables.

Separately, several scholars have used copulas to model different hydrologic processes. One of the first to make an interesting contribution were Kelly and Krzysztofowicz (1997), who used a method related to copulas in hydrology. Actual copulas were introduced, among others, by Salvadori and De Michele (2002). An early overview of different uses of copulas in hydrology has been made by Renard and Lang (2007). They only acknowledge field significance determination, regional risk analysis, Discharge-Duration-Frequency models and regional frequency analysis, so no complete hydrologic (rainfall-runoff) model. Salvadori and De Michele (2007) also give an overview in the same year, where they acknowledge the use of copulas for partial processes in the hydrologic system, such as return periods of bi- and trivariate events, similarly to work they did before (Salvadori and De Michele, 2004).

Pioneering work in using Bayesian networks as hydrologic models has been done by Molina et al. (2005), of which the complete emergency decision support model featured a rudimental hydrologic model. Then, Paprotny and Morales-Nápoles (2017) made a basic Bayesian network as a hydrologic model with data from more than 1800 European catchments and as a target value the annual maximum discharge per catchment. As input were the maximum precipitation event and several catchment characteristics, such as the area, steepness and the number of lakes and marshes used. This research gave some promising results for Bayesian networks as hydrologic models, but did still not achieve the accuracy of results of purposefully built conventional hydrologic models. The main purpose of the models was to give an educated estimate for annual runoff maximums. Couasnon (2017); Couasnon et al. (2018) built upon this research. They performed similar research as Paprotny and Morales-Nápoles (2017), but then in the contiguous United States. Sanjaya (2018) did a closely related case study in Java, in which a Bayesian network gave poor results for this specific case, which was equally true for the case study of Torres Alves (2018) in Ecuador. Nasr et al. (2018) use a similar method, but more successfully. In this research, the case study was in the Magdalena-Cauca Basin in Colombia.

1.3. Overview research goals and design

This thesis builds upon the research presented in the previous section and has as a contribution that the method is for the first time comprehensively implemented for (1.) a single catchment, (2.) for the first time in a Dutch catchment and (3.) for the first time in a lowland, partially managed, catchment. For this specific type of catchment, an optimal network layout is sought after and the different copulas are compared for this case. Furthermore, a whole new marginal probability distribution is introduced (Section 6.2) and another fit function for the copula is implemented (Section 6.3). This is done to be able to create a model that is less directly related to the data it is trained with and is able to extrapolate better, to create a model that can predict high discharge peaks. Moreover, to make a more optimal model, there is still more requirement for extrapolation. This is handled via a new shift function in Section 6.6. Then, to further optimise the model and determine its performance when parameters change, a comprehensive parameter sensitivity analysis with several goodness-of-fit methods is executed for this case study (Chapter 5). Finally, another novelty is that a BN model is benchmarked against a SOBEK model (Section 9.2.2), a neural network (Section 9.2.3), and a multiple linear regression model (Section 9.2.4). The majority of the calculations is done with the Python package `copulabayesnet`[2], which has been written for this thesis.

In this research, the theory of the method is introduced to start off with because data and case study decisions are base upon this method. First, the concept of copulas is introduced in Section 2.2. These are essential in connecting correlated parameters in non-parametric Bayesian networks. Several types of copulas are introduced, fit methods are explained and the step to the Bayesian networks is made in Appendix B.1 and Section 2.6.

Afterwards, the case study is introduced in Chapter 3. The catchment is located in the area managed by the water board Waterschap Drents Overijsselse Delta (WDODelta), in the Dutch provinces of Drenthe and Overijssel, adjoined to the province of Friesland. For this research, WDODelta has kindly made all their data available. This is a very broad and useful dataset, and this catchment is therefore highly suitable for a case study. Alongside the WDODelta data, other sources of data are used. The variables are explained in this chapter and the influence on the hydrologic system is discussed. All of the data is filtered and tested with a water balance, which is also featured in this chapter. To test parameters to be used, a first model is introduced. Finally, the metric of performance of any model is introduced in Section 4.2.

To select the optimal type of copulas to use in the model, a number of goodness-of-fit tests are introduced and implemented in Chapter 5. The performance of the copulas is also calculated, and possible weak points of the copulas are pointed out.

A copula has uniform marginals, which calls for a function to project the data from its original shape to uniform values. This is a cumulative distribution function (CDF) type of equation, of which three types are introduced in Chapter 6: two probability distributions are fitted to the data and one is a step function directly acquired from the data. All parameters are optimised and the optimal function for the model is selected.

In Chapter 7, all other parameters within the model are thoroughly tested, and the parameter that gives the best fit to predict the MMDAD is chosen. Moreover, an analysis is made of the variables to use, and how to combine these in the final model, in relationship to what gives the best performance and the most insight into the hydrologic processes in the catchment.

The results of the final model are presented in Chapter 9. In the same chapter, the SOBEK model, the neural network and multiple linear regression model, against which the model proposed in this research is benchmarked, are introduced. Afterwards, the results are discussed and put into context in Chapter 10 and a conclusion of this research is drawn, of which some recommendations follow, both in Chapter 11.

[2] `pip install copulabayesnet`, available on `https://github.com/SjoerdGn/copulabayesnet`, see also Appendix G.4.

2

Copulas and non-parametric Bayesian networks

2.1. Bayesian network

A Bayesian network (BN) is a probabilistic model that is graphically represented in the form of a directed, acyclic graph (DAG). These are graphs in which the edges also have an orientation (denoted with an arrow) and in which it is not possible to return to the same node following the arrows (Hanea et al., 2006). A BN is based around a number of variables (the nodes) and the conditional correlations between them. The method offers great flexibility in handling variables from different sources and uses probability distributions of variables throughout the network, also when the network is conditioned (i.e. certain variables are fixed). In this research, it is implemented on a relatively flat, partly managed, lowland catchment. It is examined how well it performs in predicting the discharge out of this catchment and how well it models other hydro-meteorological processes.

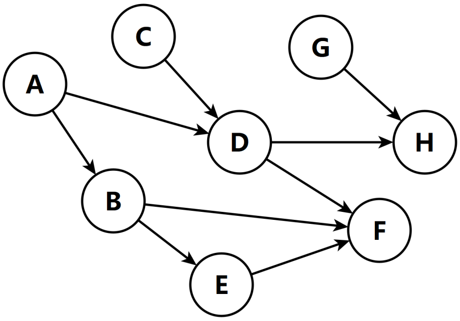


Figure 2.1: Example of a DAG. The connections have a defined direction and no cycles can be made following the direction of the arrows. In the case of a BN, the nodes are variables and the arcs are conditional distributions. See also Appendix A for a more in-depth explanation of the BN layout.

Most variables that can be measured in a catchment area, are continuous variables. A type of BN that handles these continuous variables, is the non-parametric Bayesian network (NPBN). This method accepts any distribution, as long as the continuous variables have invertable distributions (Hanea et al., 2015). This flexibility is the reason why this type of BN is used in this research. The connections, which are conditional probability distributions, are represented by multivariate probability distributions with uniform marginals, called copulas, in an NPBN. These were introduced by Sklar (1959).

2.2. Sklar's theorem

A model to connect multiple empirical distributions, is the copula. This is first defined by Sklar (1959), in what is now called Sklar's theorem. This states that every multivariate distribution with d variables $(v_1, \ldots, v_d)$, with the marginals $F_1(v_1), \ldots, F_d(v_d)$, can be written as:

$$F(v_1, \ldots, v_d) = C\left(F_1(v_1), \ldots, F_d(v_d)\right). \tag{2.1}$$

The function $C(\cdot, \ldots, \cdot)$ is the cumulative representation of the copula. For the probability density case, using the chain rule, the distribution is now defined as

$$f(v_1, \ldots, v_d) = c_{1,\ldots,d}\left(F_1(v_1), \ldots, F_d(v_d)\right) \cdot \prod_{i=1}^{d} f_i(v_i), \tag{2.2}$$

(Aas et al., 2009). Acquiring these uniform marginals from a variable, means that $F(\cdot)$ is the cumulative distribution function (CDF) of this variable. The uniform[1] marginals ($u \in [0, 1]$ and equally distributed) are also defined as u:

$$u = F(v). \tag{2.3}$$

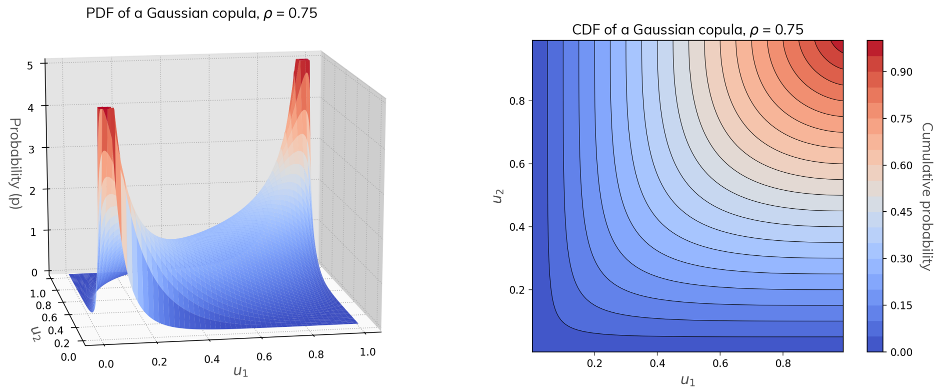
This gives that the probability density of a value of a variable v is distributed as $f(v)$. Three types of functions are proposed for $F(v)$ in Chapter 6. An example of the visual representation copula of a copula, in this case Gaussian or normal copula, can be seen in Figure 2.2.

Copulas have the advantages that any probability distribution can be used to relate variables, as long as the CDF is defined. This makes for a great method in a complicated system where not all complex relations between variables are easily expressible in other mathematical forms. In Figure 2.3, a visual example can be found of the construction of a bivariate copula.

2.3. Types of copula

Over the years, many different copula types have been proposed, which are different formulas that can be used to describe Equations (2.1) and (2.2). In this thesis, the Gaussian copula is used, which assumes a multivariate normal distribution between the inverse standard normal function of the marginals. However, there is also the family of Archimedean copulas, of which many have a different dependence in each opposite corner. This is called tail dependence.

[1] In this thesis values that originate from the uniform distribution between 0 and 1 are called *uniform values*. However, is the case of a conditioned copula, these values do not follow a uniform distribution anymore. But as they have a similar role in association with a copula, they ware still called uniform values. Any value u that can be converted to a value of a certain variable with $F_i^{-1}(u)$ is called a uniform value in this thesis.



(a) Probability density function (PDF) ($c(u_1, u_2)$) (b) CDF ($C(u_1, u_2)$)

Figure 2.2: Examples of a bivariate Gaussian copula. The CDF is plotted in a 2-dimensional colourmap, because this gives a better comparison between the different copulas. Note that the colour scale is different for the PDF and the CDF and that the 3-dimensional PDF graph is cutoff at $p = 5$.

The main purpose of testing these copulas is to test the assumption that the Gaussian copula is sufficient for all variable combinations. Some variable combinations potentially have tail dependence. If Archimedean copulas give better fits to these variables, it suggests that the Gaussian copula is imperfect. Other copula types, such as copulas based on mixture distributions, will not be analysed in this research.

2.3.1. GAUSSIAN

Gaussian copulas are symmetrical copulas that are based on the Gaussian or normal distribution. For correlation matrix R ($\mathbb{R}^d$), which represents the correlation of all variable pairs (see Section 2.6.2), the CDF of the copula (C) and the PDF (c) are defined by as

$$C_R^{\mathrm{Ga}}(u) = \Phi_R\left(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)\right), \tag{2.4a}$$

$$C_R^{\mathrm{Ga}}(u) = \frac{1}{\sqrt{|R|}} \exp\left(-\frac{1}{2}\mathbb{Q}^T \cdot \left(R^{-1} - I\right) \cdot \mathbb{Q}\right), \tag{2.4b}$$

where
$|\cdot|$ denotes the determinant,
I is the identity matrix,
$\mathbb{Q}$ is $\begin{pmatrix} \Phi^{-1}(u_1) \\ \vdots \\ \Phi^{-1}(u_d) \end{pmatrix}$
and $\Phi^{-1}(\cdot)$ is the inverse CDF of the standard normal distribution (Arbenz, 2013).

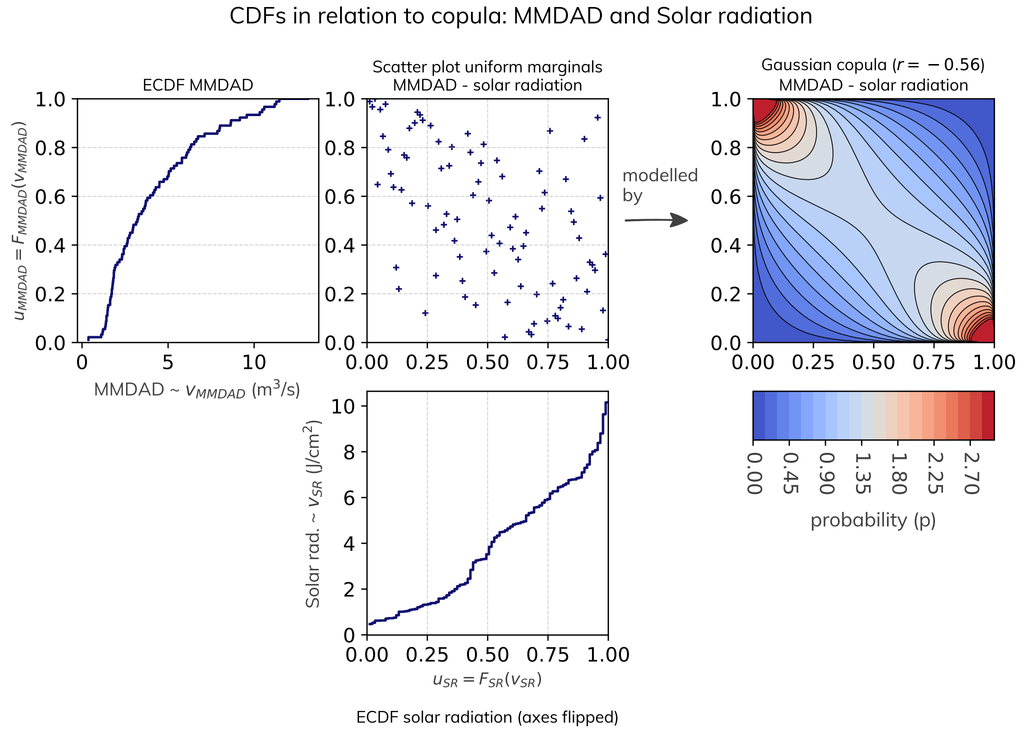


Figure 2.3: Example two variables, the monthly maximum daily average discharge (MMDAD), see Section 3.3.2, and the solar radiation (see Section 3.3.3), that are converted to uniform values via Equation (2.3). In this case, the empirical cumulative distribution function (ECDF) is used. In Chapter 6, other versions of the CDF are introduced as well. Note that for the solar radiation variable, the axes are flipped in comparison to the common display of a CDF function, to share the axis with the graph above and project the values to the scatter plot (similarly to an exceedance probability graph). The two sets of uniformly distributed values form the uniform marginals of a copula, of which the PDF is shown in this figure.

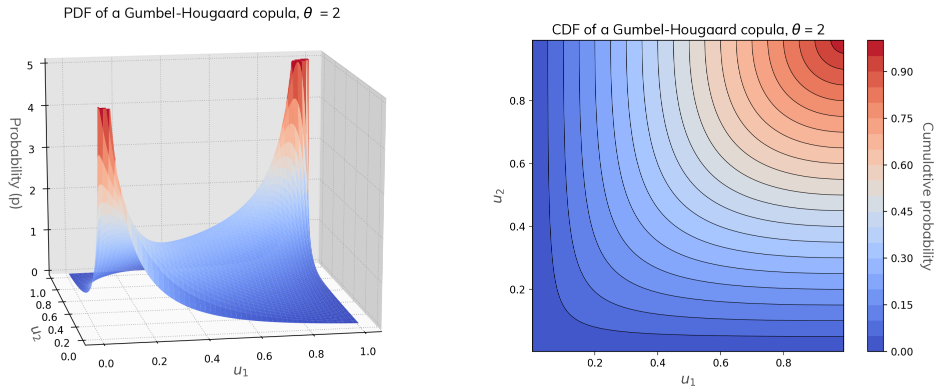
2.3.2. Archimedean copulas

Archimedean copulas generally do not have a straightforward multivariate (more than 2 variables) version. Besides some complicated multivariate extensions, other options respectively make use of vine-copulas (see Appendix B.1) or meta-elliptical copulas (Aas et al., 2009).

Archimedean copulas are characterised by the generator that each copula type has and that they have a straightforward closed form for the bivariate case. Nelsen (2006) recognises 22 types of (1-parameter) Archimedean copulas, wich are all initially defined for the bivariate case. The parameter that defines the shape of the copula is denoted with θ. In this research, only a selection of these copulas are considered for the case study model.

Gumbel-Hougaard copula

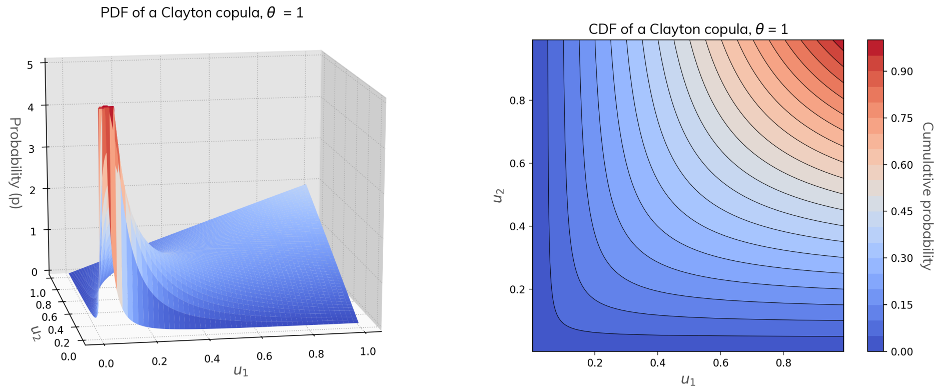
Gumbel-Hougaard copulas have upper tail dependence: this means they have a higher dependence in the north east corner in contrast to the south west corner. See Figure 2.4.



(a) PDF ($c(u_1, u_2)$)

(b) CDF ($C(u_1, u_2)$)

Figure 2.4: Examples of a bivariate Gumbel-Hougaard copula.



(a) PDF ($c(u_1, u_2)$)

(b) CDF ($C(u_1, u_2)$)

Figure 2.5: Examples a bivariate Clayton copula.

$$C_{\theta}^{\mathrm{GH}}(u_1, u_2) = \exp\left(-\left[(-\ln u_1)^{\theta} + (-\ln u_2)^{\theta}\right]^{1/\theta}\right), \quad \text{for} \quad \theta \in [1, \infty). \tag{2.5}$$

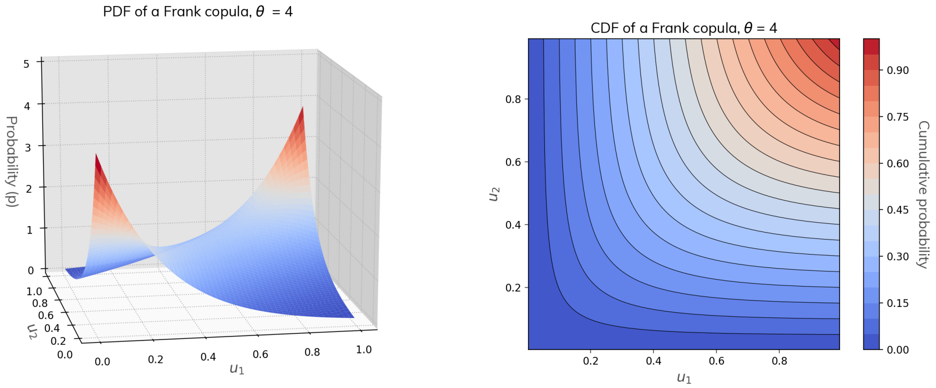
CLAYTON COPULA

Clayton copulas have a high tail dependence in the south west corner. See Figure 2.5.

$$C_{\theta}^{\mathrm{Cl}} = \left[\max\left(u_1^{-\theta} + u_2^{-\theta} - 1, 0\right)\right]^{-1/\theta}, \quad \text{for} \quad \theta \in [-1, \infty)\backslash\{0\}. \tag{2.6}$$

Frank copula

Frank copulas are similar to their Gaussian counterpart and have no tail dependence (Embrechts et al., 2001). See Figure 2.6.



(a) PDF ($c(u_1, u_2)$)

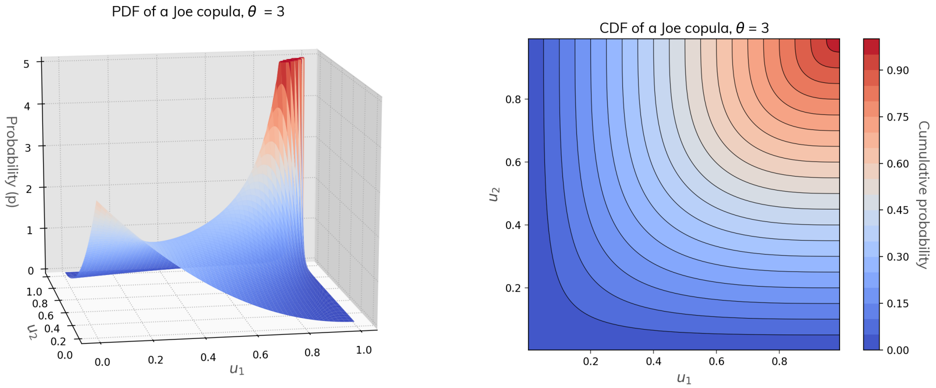
(b) CDF ($C(u_1, u_2)$)

Figure 2.6: Examples of a bivariate Frank copula.

$$C_\theta^{\text{Fr}} = -\frac{1}{\theta} \ln\left(1 + \frac{(e^{-\theta u_1} - 1)(e^{-\theta u_2} - 1)}{e^{-\theta} - 1}\right), \quad \text{for} \quad \theta \in (-\infty, \infty) \setminus \{0\}. \tag{2.7}$$

Joe copula

Joe copulas have a high tail dependence in the north east corner. See Figure 2.7.



(a) PDF ($c(u_1, u_2)$)

(b) CDF ($C(u_1, u_2)$)

Figure 2.7: Examples of a bivariate Joe copula.

$$C_{\theta}^{\mathrm{J}} = 1 - \left((1-u_1)^{\theta} + (1-u_2)^{\theta} - (1-u_1)^{\theta} \cdot (1-u_2)^{\theta}\right)^{1/\theta}, \quad \text{for} \quad \theta \in [1, \infty). \tag{2.8}$$

NEGATIVELY CORRELATED VARIABLES

For a number of Archimedean copulas, it is not possible to create a copula with a negative correlation. Therefore, in this thesis, in the case of negative correlation, one of the variables is rotated around 0 (times -1, this is equal to $u_{\mathrm{r}} = 1 - u$). It does matter, of course, which of the variables is rotated. However, as the main goal of Chapter 5 is to test for the assumption that Gaussian copulas are optimal to use for this case study, only one of the variables is rotated. This is sufficient for this research because Clayton and Joe copulas are almost the opposite of each other. In addition, the Gumbel and Clayton copula also model an opposite tail dependence. Therefore, the effect of determining whether the Gaussian copula is the optimal, is already achieved by rotating one variable.

2.4. FIT A COPULA TO DATA

In order to have the copula describe the multivariate distribution of a number of variables well, the copula has to be fitted to the data. A number of methods is available to do so. The preferred method depends on preference and the type of copula which is used. This section discusses two methods and argues which method needs to be used in what occasion.

2.4.1. MAXIMUM LIKELIHOOD ESTIMATION

A likelihood function is intuitively defined as follows: the likelihood of drawing this empirical combination, given the copula of choice with its parameter ϑ. This is a useful function as it takes all values of the variable into consideration.

CANONICAL MAXIMUM LIKELIHOOD ESTIMATION

The Canonical Maximum Likelihood Estimation (CMLE) is the maximum likelihood function that is made for copulas, and is therefore used as the preferred function to fit most of the copulas. In this research, the log-likelihood is taken. This means that the logarithm of the copula values is taken. The maximum likely parameter ϑ ($\hat{\vartheta}$) is defined as follows (SAS Institute, 2017):

$$\hat{\vartheta} = \underset{\vartheta \in \Theta}{\arg\max} \sum_{i=1}^{m} \log c_{\vartheta}(u_{i,1}, \ldots, u_{i,d}), \tag{2.9}$$

where m is the number of variables in the training set.

2.4.2. DIRECTLY FROM CORRELATION

For Gaussian copulas, the parameter that has to be fitted is the correlation coefficient, or a correlation matrix R (see Equation (2.4)). This coefficient can be fitted to the data directly with one of the previous two methods. However, it is also possible to calculate the correlation coefficient directly. Three different types of correlation coefficients are discussed in this thesis. An example of a likelihood function for a copula is to calculate the probability of each combination of uniform values in the PDF of the copula and sum this. When the logarithm of each of these numbers is taken, the method tempers the high peaks that a copula can have, such that the fit is not dominated by the errors in the corners. To find the maximum likelihood, a Nelder-Mead algorithm can be implemented, for example.

Pearson correlation coefficient

Pearson correlation is based upon how much the variables show linear correlation and is the most widely used. Its coefficient is defined according to Equation (2.10).

$$\rho_{\mathrm{p}}({V'}_1, {V'}_2) = \frac{\mathrm{cov}({V'}_1, {V'}_2)}{\sigma_{{V'}_1}\sigma_{{V'}_2}}, \tag{2.10}$$

where $\mathrm{cov}({V'}_1, {V'}_2)$ is the covariance between ${V'}_1$ and ${V'}_2$ and σ denotes the standard deviation of a variable. A Pearson correlation of 1 is defined as when parameter ${V'}_2$ is a direct positive linear translation of parameter ${V'}_1$.

Spearman's rank correlation coefficient

Spearman's rank correlation coefficient is defined according to Equation (2.11).

$$r_{\mathrm{s}}({V'}_1, {V'}_2) = \rho_{\mathrm{p}}(\mathrm{rg}_{{V'}_1}, \mathrm{rg}_{{V'}_2}) = \frac{\mathrm{cov}(\mathrm{rg}_{{V'}_1}, \mathrm{rg}_{{V'}_2})}{\sigma_{\mathrm{rg}_{{V'}_1}}\sigma_{\mathrm{rg}_{{V'}_2}}}, \tag{2.11}$$

where the only difference with Equation (2.10) is that the covariance is not taken from the values that make up the variable, but from the rank of the values (rg). The rank is the number of the value in the list, if all values would have been order from low to high. This has as an advantage that any two variables that are not correlated linearly, but still strictly increasing, still have a correlation coefficient of 1. This is exactly what is done by creating a copula, as the CDF $F(\cdot)$ also orders all values in a rank.

Normal rank correlation

Thirdly, there is the normal rank correlation, which is defined as follows (Hanea and Harrington, 2009):

$$r_{\mathrm{s,norm}}({U'}_1, {U'}_2) = \frac{6}{\pi}\arcsin\left(\frac{\rho_{\mathrm{p}}\left(\Phi^{-1}({U'}_1), \Phi^{-1}({U'}_2)\right)}{2}\right), \tag{2.12}$$

where U' is the vector of the uniform variables (also defined in Equation (2.13b)). The normal rank correlation is the rank correlation between variables ${V'}_1$ and ${V'}_2$ if its relation were perfectly defined by the Gaussian copula. Therefore, this is the correlation coefficient that is used for Gaussian copula models. However, the R-matrix for a Bayesian network is not necessarily constructed of the $r_{\mathrm{s,norm}}$ for each variable pair. This topic is covered in Section 2.6.2.

2.5. Application of copulas in non-parametric Bayesian networks

In a NPBN, the multivariate probability distribution is modelled by a copula or copulas. There is a number of approaches to apply these copulas in a BN, such that the network can be conditioned. In this section, three of these are discussed and the preferred method is selected for this thesis:

- Model the network as a single Gaussian copula. The Gaussian copula has a clear definition for multivariate (more than 2 variables) usage. Because its parameter (R) is a matrix with all the variable combinations, it is possible to implement different parameters per combination (see Equation (2.4)). Moreover, the Gaussian copula also offers the possibility to use the multivariate normal method, which offers great flexibility in conditioning (see Section 2.6). Therefore, this method is chosen for this research. The assumption that this type of copula fits the data sufficiently well, is verified in Chapter 5.

- Model the BN as one single Archimedean multivariate copula. This is unfavourable as Archimedean copulas are not all defined straightforward in the multivariate form and many have specific tail dependence, which is likely to not perform well for all variable combinations. Moreover, using a single Archimedean copula requires a single fit parameter θ to be used for all variable pairs, which highly likely will not suit all combinations.

- To make the implementation of different Archimedean copulas possible in a BN, complex conjunction of copulas must be implemented. The principal method to do this, is the vine-copula. This method combines bivariate copulas with so-called vines. However, when using for example 8 variables, such as proposed in Table 4.1, this method requires 20160 different vines that have to be calculated (see Equation (B.3) in the appendix). This is highly unfavourable. See Appendix B.1 for a more in-depth coverage of vine-copulas.

2.6. MULTIVARIATE NORMAL METHOD

The (multidimensional) Gaussian copula is defined by the multivariate CDF of the inverse standard normal CDFs of the variables (see Equation (2.4a)). Conditioning of this copula can therefore be done by using the multivariate normal (MVN) distribution (Φ_R) directly. This makes it more straightforward to calculate a network which is not conditioned on all variables (Hanea et al., 2006).

2.6.1. PROCESS

For a set of values (one value per variable) $V = [v_1, \ldots, v_d]^T$, $U = [u_1, \ldots, u_d]^T$ is defined as a vector with the uniform values u of the corresponding values v of these parameters:

$$u_i = F_i(v_i), \tag{2.13a}$$

$$U = [F_1(v_1), \ldots, F_d(v_d)]^T. \tag{2.13b}$$

These uniform values are then transformed to standard normal values[2] with the inverse cumulative distribution function of the standard normal distribution, Φ (i.e. $\mathcal{N}(0,1)$):

$$s_i = \Phi^{-1}(u_i), \tag{2.14a}$$

$$S = \Phi^{-1}(U). \tag{2.14b}$$

[2] As a similar generalised term as uniform values, standard normal values are any values s which can be converted to uniform values with $\Phi^{-1}(u)$.

Now a conditional value v_k given the conditioning (fixed) values $V_c = [v_{c,1}, \ldots, v_{c,z}]^T$ is distributed according to G_k:

$$v_k|V_c \sim G_k(S_c) = F_k^{-1}(\Phi(s_k|S_c)), \tag{2.15}$$

where S_c is defined by Equation (2.14b), $s_{i,\ldots,z}$ are the values of the z conditioning variables, s_k follows from Equation (2.14a) and the conditioning is done with:

$$s_i|S_c \sim \mathcal{N}(\bar{\mu}(U_c), \bar{R}(U_c)). \tag{2.16}$$

In this equation, $\bar{\mu}(U_c)$ is defined according to Equation (2.18) and $\bar{R}(U_c)$ according to Equation (2.19) in Section 2.6.3 (Hanea et al., 2006). Figure 2.8 contains a visual overview of the multivariate normal method.

2.6.2. CORRELATION MATRIX

The correlation matrix R of the BN is based on the partial correlations between all variables (Hanea et al., 2006). For each position in the correlation matrix, the generic correlation coefficient r for the variables V'_1 and V'_2, given by the other variables (denoted in the subscripts), is defined as:

$$r_{12;3,\ldots,d} = \frac{r_{12;4,\ldots,d} - r_{13;4,\ldots,d} \cdot r_{23;4,\ldots,d}}{\sqrt{(1 - r^2_{13;4,\ldots,d}) \cdot (1 - r^2_{23;4,\ldots,d})}}. \tag{2.17}$$

In this case, the correlation coefficient r that is used is the normal rank correlation $r_{\mathrm{s,norm}}$ (Equation (2.12)). Now, because the correlation is dependent on other partial correlations coefficients, this becomes a recursive formula, for which the order of correlations between parameters matters. In practical terms, the rank order of arrows in the BN matters to the values in this matrix, and therefore the outcome of this model. In this research, the matrix R is acquired by putting the model into the commercially available software Uninet. Uninet enables ordering the variables manually.

2.6.3. CONDITIONAL MVN PARAMETERS

As Gaussian copulas are used, the uniform marginals can be rewritten as a Gaussian distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$, that is $\Phi^{-1}(u)$. The conditional multidimensional Gaussian distribution is then defined with mean $\bar{\mu}$, as follows:

$$\bar{\mu} = \mu_k + R_{kc}R_{cc}^{-1}\left(A - \mu_c\right), \tag{2.18}$$

and as correlation matrix $\bar{R}$:

$$\bar{R} = R_{kk} - R_{kc}R_{cc}^{-1}R_{ck}, \tag{2.19}$$

where $\mu_k = \left[0\ldots0\right]^T$ with length $(d - z)$, where d is the number of variables in the multivariate distribution and the number of nodes in the BN, and z is the number of variables which are fixed, such that $(d - z)$ is equal to the number of unfixed, conditional variables, $\mu_c = \left[0\ldots0\right]^T$ with length z, as the means of the initial distribution are 0, and $A = \left[a_1, \ldots, a_z\right]^T$ is made up of the z standard normal values of the conditioning variables:

$$a_i = \Phi^{-1}(F_i(v_{c,i})) = \Phi^{-1}(u_{c,i}). \tag{2.20}$$

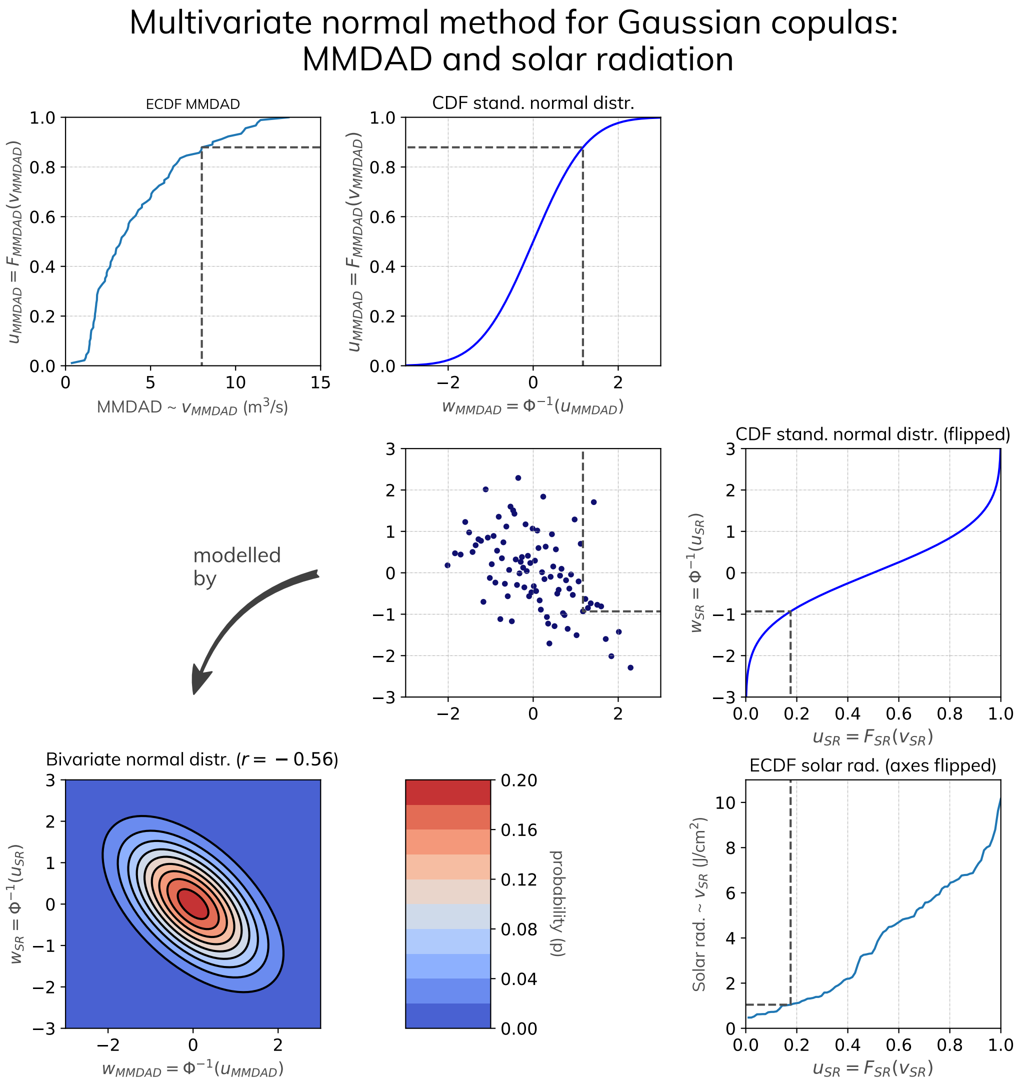


Figure 2.8: Example two variables, the MMDAD (see Section 3.3.2) and the solar radiation (see Section 3.3.3), that are converted to uniform values via Equation (2.3) and then converted into standard normal values via the CDF of the standard normal distribution. In this case, the empirical CDF is used. In Chapter 6, other versions of the CDF are introduced as well. The dark grey striped line represents an example timestep with $v_{MMDAD} = 8.0$ and $v_{SR} = 1.0$, which is plotted on the multivariate normal space as (1.17, -0.93) (in the plot visible as (-0.93, 1.17) because the axes flip). The combination of the standard values of both variables is modelled by the multivariate normal distribution, of which the PDF is shown in this figure.

Furthermore, the subsets of R are defined as follows:

$$R = \begin{bmatrix} R_{kk} & R_{kc} \\ R_{ck} & R_{cc} \end{bmatrix} \text{ with sizes } \begin{bmatrix} (d-z)\times(d-z) & (d-z)\times z \\ z\times(d-z) & z\times z \end{bmatrix}, \tag{2.21}$$

where R is the correlation matrix of the network. The sub matrices $R_{kk},\ldots,R_{cc}$ consist of the indices of the fixed variables when a size z is given and the corresponding subscript is k, and of the indices of the variable with a conditioned distribution when a size $(d-z)$ is given and the corresponding subscript is c (Helwig, 2017). Thus, the locations of the subsets of R are not necessarily in the same position as shown in Equation (2.21) . This way, a network can be conditioned with the R of the network (see Section 2.4.2) and any number of conditioning values.

2.7. Random sampling

For the verification of the copula (Section 5.3) and the conditioning of a copula (Section 2.8), the copula has to be sampled randomly. According to Embrechts et al. (2001), n random samples of a Gaussian copula $[u_1,\ldots,u_n]^T \sim C_R^{Ga}$ can be generated as follows:

1. Find the lower triangular matrix L of the Cholesky decomposition of the correlation matrix R.
2. Draw n independent random values $W=[w_1,\ldots,w_n]^T$ from $\mathcal{N}(0,1)$.
3. Set $S=LW$, with $S=s_1,\ldots,s_n$
4. Now $u=\Phi(S)$ with $i=1,\ldots,n$.

Calculating n random samples of a multivariate normal distribution is very similar (Gentle, 2009):

1. Find the lower triangular matrix L of the Cholesky decomposition of the conditioned correlation matrix $\bar{R}$.
2. Draw n independent random values $W=[w_1,\ldots,w_n]^T$ from $\mathcal{N}(0,1)$.
3. Set $S=LW+\bar{\mu}$, with $S=s_1,\ldots,s_n$

Sampling from an Archimedean copula is done by using its generators and its Laplace-Stieltjes transform, see for example Hofert (2008).

2.8. Calculating conditional distributions

To use information that is known in a catchment to update the probability distribution of other variables, the copula has to be conditioned. Unconditioned, v is distributed as follows:

$$v_u = F^{-1}(u_u) \quad \text{with} \quad u_u \sim \mathcal{U}(0,1), \tag{2.22}$$

where the subscript $_u$ denotes that it is unconditioned and $\mathcal{U}(0,1)$ is the uniform distribution between 0 and 1. However, when a copula is conditioned, u is no longer distributed by the PDF definition of the copula $\mathcal{U}(0,1)$, but by $c(u_1,\ldots,u_c,\ldots u_n)$, where u_c is the conditioned variables on all of the other variables: these are now represented by a single number.

$$v_c = F^{-1}(u_c) \quad \text{with} \quad u_c \sim c(u_1,\ldots,u_c,\ldots u_n). \tag{2.23}$$

In this section, a closer look is taken at the bivariate case, as an illustration. The two-dimensional copula function is defined as follows from Equation (2.2):

$$f(v_1, v_2) = c(u_1, u_2) \cdot f(v_1) \cdot f(v_2), \tag{2.24}$$

which results in, given Bayes' theorem:

$$f(v_1|v_2) = \frac{f(v_1, v_2)}{f(v_2)} = c(u_1, u_2) \cdot f(v_1), \tag{2.25}$$

in which $c(u_1, u_2)$ is the PDF of the copula function (for example the Gaussian copula as given by (2.4b)), and f_1 and f_2 are the PDFs of the marginals. As $f(v_1)$ is not defined well, this function cannot be calculated directly. If, for example, an expected value is required, a stepwise approach needs to be taken. This can be done in two ways:

2.8.1. BY SAMPLING REGULARLY

To directly take values from the copula function, the uniform marginal needs to be sampled regularly, as taking the variable regularly gives an irregular sample at the copula, which distorts the probability distribution within the copula. The method uses the copula probability as a factor of the values v. The following formulas calculate the expected value with this method. Take n values of u with a regular interval: $U_g = [u_{(1,1)}, \ldots, u_{(1,n)}]^T = [0, 1/n, 2/n, \ldots, (n-1)/n, 1]^T$. Now, the updated distribution becomes:

$$\mathbf{E}f(v_1|v_2) \approx \frac{1}{n} F_1^{-1}(U_g)^T c(U, u_2) = \frac{1}{n} \sum_{i=1}^{n} F^{-1}(u_{(1,i)}) \cdot c(u_{(1,i)}, u_2). \tag{2.26}$$

For the multivariate normal method, the following holds, when 0 and 1 are removed from U.

$$\mathbf{E}f(v_1|v_2) \approx \frac{1}{n} F_1^{-1}(U_g)^T \left(\bar{r} \cdot \phi(\Phi^{-1}(U)) + \bar{\mu}\right) = \frac{1}{n} \sum_{i=1}^{n} F^{-1}(u_{(1,i)}) \cdot \left(\bar{r} \cdot \phi(\Phi^{-1}(u_{(1,i)}) + \bar{\mu}\right), \tag{2.27}$$

where $\bar{\mu}$ follows from Equation (2.18) and $\bar{r}$ is the corresponding correlation coefficient for that variable pair from Equation (2.19).

2.8.2. BY SAMPLING RANDOMLY

It is also possible to sample the copula randomly, via Section 2.7. If all the the values from the random sampled values from the copula are taken that are between $u_2 - \epsilon$ and $u_2 + \epsilon$, with ϵ as a small value, only values are used that are approximately on the line of the conditioned variable $v_2 \to u_2$. These values are denoted as $U_r = [u_{(r,1)}, \ldots, u_{(r,n)}]^T$. Here, $F_1^{-1}(U_r)$ can be used as a approximation of the conditional distribution.

$$\mathbf{E}f(v_1|v_2) \approx \frac{1}{n} \sum_{i=1}^{n} F_1^{-1}(u_{r,i}). \tag{2.28}$$

For the multivariate normal case, the final conditional distribution can be sampled as well, as it is just another normal distribution (Section 2.7 with the updated R matrix). Therefore, no ϵ is needed. The sampled the values are $S_r = [s_{r,1}, \ldots, s_{r,n}]^T$.

$$\mathbf{E}f(v_1|v_2) \approx \frac{1}{n} \sum_{i=1}^{n} F_1^{-1}(\Phi(s_{r,i})). \tag{2.29}$$

When no summation is made, this method allows for easily calculation a median and quantiles of the data, to be used in a confidence interval, for example. That is why this method is used in the rest of this thesis.

3
Case Study

3.1. Selecting case study

This research focuses on a single catchment, as the goal of this research is predominantly creating an optimal model. For this single catchment, all of the data is examined and the processes are described extensively, to thoroughly study the model and its workings in the catchment. To do research that is also partly applicable in other lowland catchments, the catchment used in this thesis should be representative for lowland catchments. Furthermore, in order to be able to construct a useful Bayesian network (BN), data that is available in this catchment, should have high enough temporal frequencies, as for every month the data is subset on several days, and should be spatially representative for the whole catchment. The data for this catchment should also cover a long temporal range, to have enough training timesteps, and have a relatively low amount of errors. Furthermore, data of various variables should be available, to be able to model different catchment processes.

However, the data that is available for the case study, should not have a extraordinarily better quality (i.e. high temporal, spatial density, long timespan, low amount of errors) than in most other lowland catchments. The reason for this, is that this will cause this research not to be representative anymore of similar catchments. Based upon this reasoning, the catchment of the Vledder, Wapserveense and Steenwijker Aa[1] is chosen.

3.2. Vledder, Wapserveense and Steenwijker Aa

The catchment to which the method is tested, is that of the Vledder and Wapserveense Aa, flowing into the Steenwijker Aa. Another important tributary is the managed canal called Nijensleker Schipsloot.

It is a partly managed river system of approximately 180 km^2 with a nature reserve in the north. Furthermore, the catchment consists of mostly agriculture and discharges into the town of Steenwijk. An overview of the catchment can be found in Figure 3.2 and Figure D.1 in the appendix, for an overview with a clear digital elevation map.

[1] The main rivers are Vledder Aa, Wapserveense Aa and Steenwijker Aa, as ’Vledder’, ’Wapserveense’ and ’Steenwijker’ are adjectives with the location.

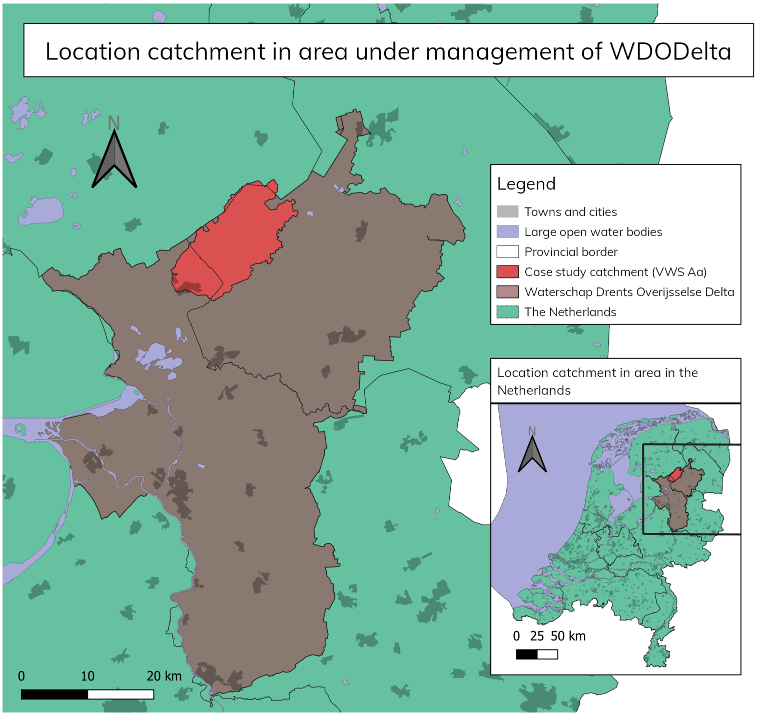

Location catchment in area under management of WDODelta
N
Legend
Towns and cities
Large open water bodies
Provincial border
Case study catchment (VWS Aa)
Waterschap Drents Overijsselse Delta
The Netherlands
Location catchment in area in the Netherlands
N
0 25 50 km
0 10 20 km

Figure 3.1: Location of the catchment in The Netherlands and in the area under management by Waterschap Drents Overijsselse Delta (WDODelta).

3.2.1. History

The larger, upstream section of the catchment, the Vledder and Wapserveense Aa coped with many floods, prior to the 1950s. Therefore, a water board called De Vledder en Wapserveense Aa, was erected in 1950 (Nieuwsblad van het Noorden, 1950). This implemented many canalisation works to make the river flow less naturally (Friese Koerier, 1953), which proved successful in battling floods (Friese Koerier, 1965). During this time, agriculture area and volume was increased in the catchment.

3.2.2. Recent works in the catchment

In 2002 and 2003, some of the agricultural area in the northern part of the catchment was converted to nature and the upstream section of the Vledder Aa was reshaped to meander again (Langendijk et al., 2014). During this time, also the Moordstuw[2] was removed, because the river now experienced higher resistance (RTV Drenthe, 2014).

The middle section of the Vledder Aa was remodeled to have a more natural flow regime in 2014, by adding artificial meanders. On top of that, the cross section of the river is made more natural, which causes more friction in the system (Langendijk et al., 2014). At the same time, 200.000 m^3 of area was created to harvest rainwater for dry spells (Zandstra, 2016). To establish

[2] *Murder weir* in English, named after an attempted murder in 1964 (RTV Drenthe, 2014)

more fish migration, several fish ladders over the whole catchment have been created, of which the biggest were instated in the period 2014-2016.

As additional research, it would be interesting to look at the influences of the works in the catchment on the discharge in relation to various variables in the catchment. In order to identify these influences, a model has to be made before and after one of these changes, that both have an abundantly high data quantity to create a representative model. This amount of data is relatively large in comparison with conceptual hydrologic models, as the non-parametric Bayesian network (NPBN) is entirely data driven. As there have been a multitude of works in the catchment, of which the implementation also took a significant time period, it is not possible to select a number of time periods between different works in the catchment that can form the basis of a reliable model. As a result, any research into influences of works in the catchment is not possible with the data used in this thesis.

3.2.3. Catchment delineation

Method

It is essential to make a good approximation of the location of the catchment border, for a number of reasons. Firstly, in order to make a water balance (see Section 3.4), the area of the catchment needs to be known. This is because some of the variables in the water balances are volumes (discharge for example) and others are lengths (such as precipitation). To compare the variables in the same units, the catchment area is required. Furthermore, it is also useful to know what variables might come into play in the catchment; for both the water balance (for example drinking water wells) as well as the BN (for example, if there are large surface water areas in the catchment, the surface water variable might play a more central role in the BN layout). Finally, of some of these variables, the location of the catchment is needed to see which measurements influences the studied catchment directly (for example, the location of soil moisture measurements inside or outside the catchment area).

The catchment has been delineated with a combination of two approaches: (i) the water level zones of the water board[3] and (ii) a watershed delineation in GRASS, an open-source software package with a broad range of hydrology tools.

i: The water level zones in The Netherlands are zones with a legal status that are put in place to give inhabitants, most importantly farmers, a degree of certainty over water levels. This is mainly useful for (partly) managed catchments with low elevations. In most cases, as also applies to the water level zones of the case study, these zones drain into one point at the edge of the zone. Therefore, it is possible to make an initial delineation out of the water level zones with these water level zones.

ii: Far upstream in the catchment in the Vledder Aa and the Tilgrup, the flow is mostly natural and freely discharging. Therefore, water levels are hardly managed and exact locations of water level zones are not so clear. This is because the exact border is not considered to be very relevant for the water board. Therefore, the choice is made to shift the border of the catchment to the border generated with GRASS in these specific areas.

Verlengde Nijensleker Schipsloot

In the north-west of the catchment, the canal called the Verlengde Nijensleker Schipsloot is positioned close to and across the catchment boundary. This is a section that has a low ele-

[3] *Peilgebieden* in Dutch

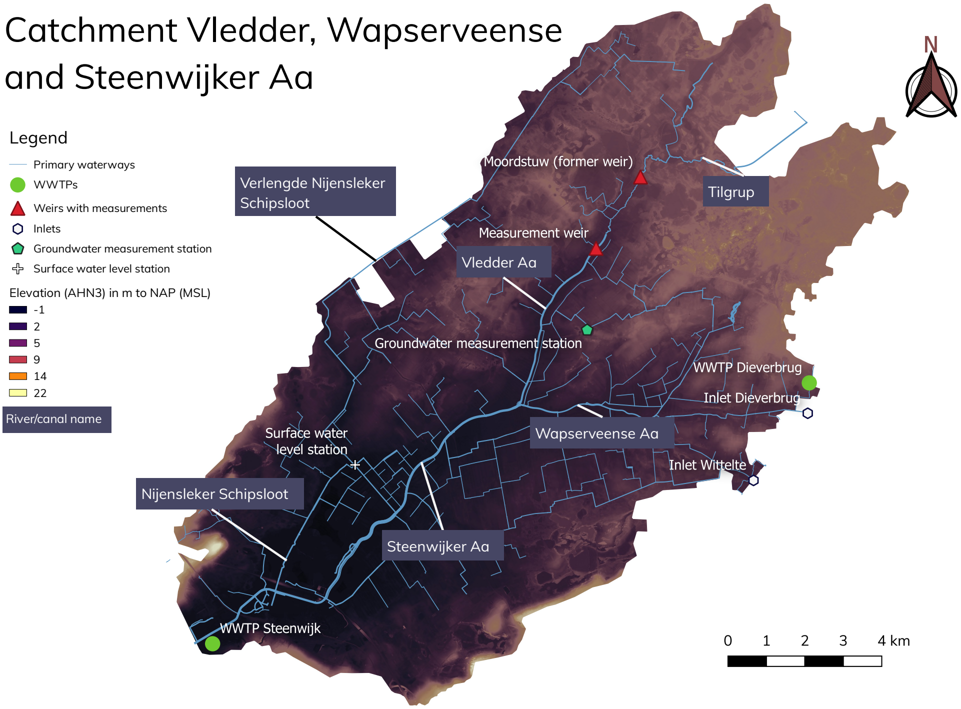


Figure 3.2: Catchment map of Vledder, Wapserveense and Steenwijker Aa, with the names of waterways and other notable items.

vation and is fairly flat, and is therefore discharging to the catchment north of the catchment of this case study. To accomplish this, there are culverts underneath this canal (conversations with Zwannie Visser (2020) and Waterschap Drents Overijsselse Delta (2019)). A small seepage discharge can be expected out of this Verlengde Nijensleker Schipsloot.

3.2.4. Waterways

The catchment consists of a partly natural river system and some agricultural canals. These are the main streams, as can be found in Figure 3.2:

- Vledder Aa. This is the river that branches out almost to the north-north-eastern border, passing the measurement weir and the former Moordstuw.

- Tilgrup. This is a highly natural tributary to the Vledder Aa that branches out to the pumping station Bosweg, where it becomes an argicultural canal upstream of the catchment.

- Wapserveense Aa. This is the largest tributary of the Vledder Aa, which is connected to the inlets at Dieverbrug and Wittelte.

- Steenwijker Aa. Somewhere downstream of the merge between the Vledder Aa and the Wapserveense Aa, the Vledder Aa changes name to the Steenwijker Aa.

- Nijensleeker Schipsloot. One of the largest canals in the area and the most managed part of the catchment. The canal passes three small pumping stations in a row and the surface water level station. It continues near the catchment border as the 'Verlengde (Elongated) Nijensleeker Schipsloot'.

Along and close to the south-eastern border of the catchment at the inlets of Dieverbrug and Wittelte is a large canal, that splits up the original larger catchment (see Figure D.2 in the appendix). It is possible that water is seeping into the catchment from a slightly higher elevated canal.

3.2.5. Geology

The catchment is relatively flat in comparison with European catchments, with elevations ranging between 0 and 25 m to NAP (approx. mean sea level) in the west, and upstream in the north east, up to 13 m NAP (Actueel Hoogtebestand Nederland, 2019). For Dutch catchment, this is actually moderately flat, as there are also catchments that have almost no inclination. The top soil consists of a wide range of soil types, but the most important ones are hummusy sand, highly mineral sand, podzolic soil and loamy soils (Wageningen UR, 2019).

3.2.6. Land use

The land use is divided as follows:

- Nature, small (production) forests and wastelands: approximately 35-40% of the total catchment area. Of this percentage, approx. 30% is a nature reserve (10.7% of the total area (Ministerie van Economische Zaken, 2019)). These areas are usually managed less and have different groundwater levels. Actual evaporation[4] is relatively low.

- Agriculture: approximately 43% of the catchment area (Ministerie van Economische Zaken, 2019). The water in the vicinity of agriculture is usually managed more and has a higher actual evaporation rate because of sprinkling and the vegetation. The most prominent crops are:

  - Pasture/grasslands (approx. 68% of the agricultural area and 29% of the total area)
  - Maïs (approx. 15% of the agricultural area and 6.5% of the total area)
  - Potatoes (approx 10% of the agricultural area and 4% of the total area)

- Buildup area: approximately 7 to 8% (Kadaster, 2019). A proportion of this area discharges to the sewer system instead of the catchment.

- Open water: 1.36% of the catchment area (Kadaster, 2019). The actual evaporation is higher in open water than on land.

[4] In this research, actual evaporation contains the actual evaporation as well as the actual transpiration that is happening.

- Other: 8-12%. These contain, among others, a military training ground, several holiday resorts and campsites, farms buildings and other housing outside built up areas and a small zoo.

3.2.7. CLIMATE

Together with the rest of The Netherlands, the catchment lies within a region with a marine climate. For its latitude, it has a relatively mild winter, as well as a mild summer. It rains year-round and seasonal differences in precipitation are low.

3.2.8. ARTIFICIAL STRUCTURES AND WATER FLUXES

The usage of copulas presumes that the variables are independent and identically distributed (IID). Artificial structures and management of water flows has the potential to make this assumption less valid. The following artifical structures are implemented in the catchment.

WEIRS

There are 22 weirs in the catchment (Zandstra, 2016). Most of these lie in the smaller tributaries and are not managed regularly or just fixed. Therefore, most of these weirs do not alter the assumption of the data being IID. However, there are five weirs that are larger and situated in the main streams. These control the water level, and therefore also the discharge to a certain extent. Because of this, the water level and also the discharge, behave less like IID variables.

PUMPING STATIONS

There are some minor, automatic pumping stations inside the catchment that are run by WDO-Delta. As they work automatically, they are regarded as not changing the presumption of IID discharge values in this thesis. Moreover, the total discharge that they contribute is relatively low. They can be found in Figure 3.2. There is a possibility that some farmers have their own pumps, but these are also presumed to be neglectable.

At one of the northern edges of the catchment, one of the tributaries to the Vledder Aa, the Tilgrup, stretches all the way to the edge of the catchment and actually continues as a canal further upstream, which connects to other canals as well. The area upstream is another (narrow) flat area where the water is managed at the border of the catchment by the pumping station Bosweg. This pumping station pumps water into the catchment, but the exact area in which water is discharged to the Tilgrup depends on water levels in that upstream subcatchment. This makes that the catchment will always be imperfectly delineated. However, data is available on the amount of water flowing into the catchment, so it is decided to position the catchment border just downstream of the pumping station and note the water flowing in. Since 2018, a larger area upstream of the catchment drains into the Vledder Aa, supported by the pumping station Bosweg (Langendijk et al., 2014).

INLETS

There are two inlets from the Drentse Hoofdvaart to the catchment that are used to make up for water shortages upstream at the east side of the catchment. The inlets are near to the villages of Dieverbrug and Wittelte respectively. See Figure 3.2 for this location. According to J. de Jong of the water board WDODelta (personal communication, December 19, 2019), these inlets are manually operated, causing a non-randomly distributed parameter in the mode. The discharge Q is calculated with the horizontal weir formula:

$$Q = 1.7 \cdot c_m b h_w^{3/2}, \tag{3.1}$$

where c_m is a friction factor, in this case it is fitted by WDODelta 1.00588, b is the width of the weir and h_w is the water level above the weir. The inlet of Dieverbrug is used in the water balance, but the one of Wittelte is not, as this inlet gave too many implausible values. Both inlets are not featured in the BN as their total contribution to the network is less than 1% and according to WDODelta, all the water is used upstream to sprinkle crops. In the water balance, however, it makes a small difference in the actual evaporation, so to be complete, it is implemented there.

SEWERAGE

The sewer system is neglected as the part of the catchment that is connected to the sewer is low and a lot ends up in the waste water treatment plant (WWTP) at Steenwijk, which discharges the water again into the Steenwijker Aa just upstream of the measurement location. A small portion ends up in in WWTP Dieverbrug, which does leave the catchment. See Figure 3.2 for the locations.

GROUNDWATER EXTRACTION

A short distance north of the catchment lies the groundwater extraction well Terwisscha. This well extracts approximately $6 \cdot 10^6$ m^3 water per year. Due to the extraction, the water level has declined up to 30 cm in some places (AdviesCommissie Schade Grondwater, 2015). This causes harm to the nature reserve, which is why it will be reduced in the future (Leeuwarder Courant, 2016).

As the groundwater extraction is a local difference and not known per month, this number is only used in the water balance. The extraction well is situated outside of the catchment. Therefore, the amount that is extracted is arbitrarily chosen as 45% of the total extraction.

There are also minor, privately owned groundwater pumping stations. The most important use is sprinkling agricultural fields. The amount of pumping is not known and as most of the water does not leave the catchment except for discharge and evaporation locally, influences of this pumping are not implemented in the model nor in the water balance.

3.3. DATA

3.3.1. OVERVIEW

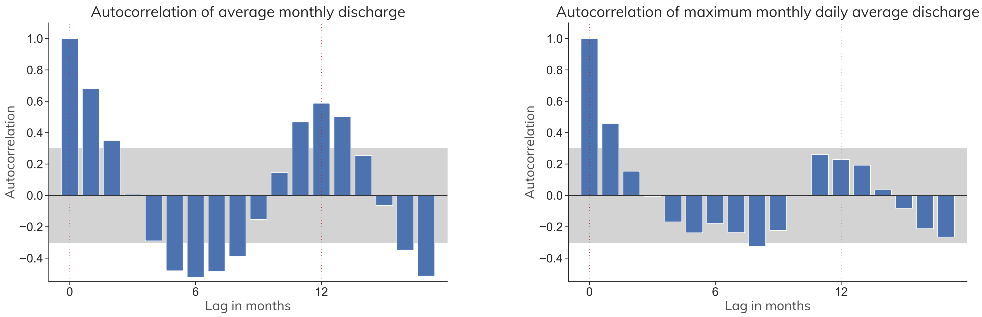
For this research, two main categories of data are used: (1.) Data to use in the BN, and (2.) Data to close the water balance to verify the data in category 1 as much as possible.

It is preferred that the data has a high spatial resolution, or has a low spatial difference for the whole catchment.

3.3.2. DISCHARGE

TARGET VARIABLE

The target variable of this research is the monthly maximum daily average discharge (MMDAD). This is an interesting variable since water boards need to make sure the land in their area does not inundate often (Rijksoverheid et al., 2003). Therefore, they should know how often floods and high discharges are to be expected, especially given the observed increase of heavy precipitation events (IPCC, 2014). The reason this variable is chosen, instead of the maximum monthly value, is that the discharge measurements fluctuate a lot. Therefore, the random error is assumed to be large. To make the influence of these errors smaller, daily average values are taken. Moreover, using this variable choice makes the studied variable more likely to be IID, which is needed for a probabilistic method, in contrast to the use of monthly averages. This can be seen in Figure 3.3, where the autocorrelation is significantly less for the MMDAD. On top of that, the correlation coefficient with one of the most important variables also improves: precipitation. For the correlation with the monthly average discharge, the Pearsons correlation coefficient is 0.31, whereas it is 0.42 when the MMDAD is used. As variables make for better predictors when the correlation between them is higher, this is favourable.



(a) Monthly average

(b) Monthly maximum daily average

Figure 3.3: Comparison autocorrelation monthly average vs monthly maximum daily average

SELECTION BASED UPON DATE MAXIMUM

As it does not make sense to use data that happened after the MMDAD event (this cannot have influenced the discharge), all the other datasets have to be subset on a period before that event.

For the first model, this period is chosen as 8 days, but it is a parameter that is optimised in Section 7.3. Over this period, the mean or the sum of the values is taken, depending on the type of variable. In essence, it does not matter if the mean or sum is chosen as the subset period is constant. It is considered to implement a formula that gives data from different days ago different weights, to get a dataset with data that influences the MMDAD the most. However, this requires knowledge about the delay time in the system and similar hydrologic parameters, which is not the purpose of this model. However, when the accuracy of the model has to be perfected more, using a function for this subset period, could be subject of further research.

All other datasets should have at least one datapoint in this period, but preferably many more, such that non-bias errors get filtered out and a good average over the time period can be obtained.

It is possible that some of these periods overlap, which is negative for the IID presumption.

Discharge measurements

The discharge measurement is taken from a measurement station just downstream of the town of Steenwijk. The discharge is measured by multiplying the average flow velocity and the cross section times a factor (see Equations (C.1) to (C.3) in the appendix for the equation that is used by the water board). The velocities are measured with side looking doppler (SLD) measurement devices.

Verification

The method is verified with an acoustic doppler current profiler (ADCP). However, this is not done regularly and for all discharge amounts. It is likely that there are errors in the discharge measurements, but these are unknown. Therefore, it is useful to take a look at the quality of the measurements, solely from the data.

Creating a histogram out of the individual measurements (per quarter of an hour), showed an approximate distribution of the measurements (Figures D.4 and D.5 in the appendix). It was decided that any discharge below -5 m^3/s (so upstream) was an outlier and was removed. The exact point of cutting these measurements does not matter so much as there was a low number of recordings below -5 m^3/s and the goal of this research is to look for the high peaks in discharges, not low values.

Moreover, there is also data available on a weir approximately 7 km upstream of the measurement location. As the discharge at this weir is solely based on a Q-h relation, which can cause large discharge measurement errors for weirs, and it is unclear whether the water over the weir is free flowing or has submerged flow, which can also cause errors in the measurements. Additionally, not all of the flow passes this weir. Therefore, this cannot be regarded as more trustworthy. The water that enters the system after this spot mostly originates from more managed sources. In spite of its untrustworthiness, the weir data has been used to point out potential measurement errors. See Figure D.6 in the appendix for the corresponding plot. The general flow at the weir is slightly lower, which is logical because it is farther upstream. Moreover, the highest peaks are significantly lower. It is possible that either of these measurements is bad at predicting these high discharges. Lastly, in the winter of 2016-2017, the discharge at the weir drops well below the measurements downstream. The only conclusion that can be drawn from this, is that this is a period of attention.

3.3.3. KNMI data

The Koninklijk Nederlands Meteorologisch Instituut (KNMI) is the Royal Dutch Meteorological Institute, which has over 100 weather stations in the Netherlands and offshore in Dutch waters. These stations are often similar to each other and measure a wide variety of atmospheric parameters. For this research, the following are used:

Temperature

The KNMI dataset contains hourly temperature measurements at 1.5 m above the ground in degrees Celcius.

Solar radiation

The KNMI stations hourly measure global radiation: the short-wave radiation falling onto a horizontal surface (PIK, 2020). The measurements are in J/cm^2. Together with the temperature, the solar radiation can serve as a proxy for the evaporation and transpiration in the catchment.

POTENTIAL EVAPORATION

The KNMI stations calculate the potential evaporation from a number of values that they measure, daily. The Makkink equation is used to acquire this number.

PRECIPITATION

The KNMI precipitation data is featured in Section 3.3.4.

SPATIALLY COMBINING KNMI DATA

The KNMI data is combined by using Thiessen polygons (Luxemburg and Coenders, 2017). These polygons are based on the Voronoi diagram. This method is chosen over the inverse distance weighting as this has an ambiguous β factor, which is not fitted for this catchment, as well as the open question of how many stations to take into account. In the case of the Vledder and Wapserveense Aa catchment, this results in two KNMI stations that are used for the measurements, that lie about 16 and 19 km outside of the catchment. For the BN and the water balance, one value is taken into account for the whole catchment: a combination of each measurement of the stations scaled by the percentage of the catchment each polygon covers. Using Thiessen Polygons, only KNMI stations 273 (Marknesse) and 279 (Hoogeveen) have any influence in the catchment, with rounded contribution factors of 0.5 and 0.5 (see Figure D.2 in the appendix).

3.3.4. PRECIPITATION

In this research, two sources of precipitation data are used.

- 2 stations of the Dutch Meteorogical Institute KNMI, which lie between 16 and 19 km outside of the catchment. However, for verification, 4 stations are used that lie up to 26 km outside of the catchment.
- 2 stations of the Water Board WDODelta, of which one lies in the centre of the catchment and the other lies close to the catchment.

The data of the KNMI is heavily verified and is therefore assumed to contain only little errors. Moreover, it was stated from WDODelta that a multitude of their stations in the northern part of the area that they manage are placed on inadequate locations, such that more or less precipitation could fall into their rain gauges than should be the case. If the errors only arise from a constant bias, the BN's accuracy would not be affected. The stations of WDODelta, however, are located closer to the catchment. Therefore, those are the preferred measuring stations if the quality is sufficiently high. To test this quality, a cumulative plot has been made with all of the stations (see Figure 3.4). Here, it can be clearly seen that the WDODelta stations measure higher rainfall numbers. The assumption is that this relatively flat area in the Netherlands should not display regional differences when averages over multiple years are taken from stations so close to one another. The chance that the types of cumulative precipitations from WDODelta are drawn out of the statistical distribution that can model the precipitation of KNMI stations as a normal distribution, is tested with a Student t-test[5]. Here, for both stations

[5] See The Editors of Encyclopaedia Britannica (2019) for an explanation of the Student t-test

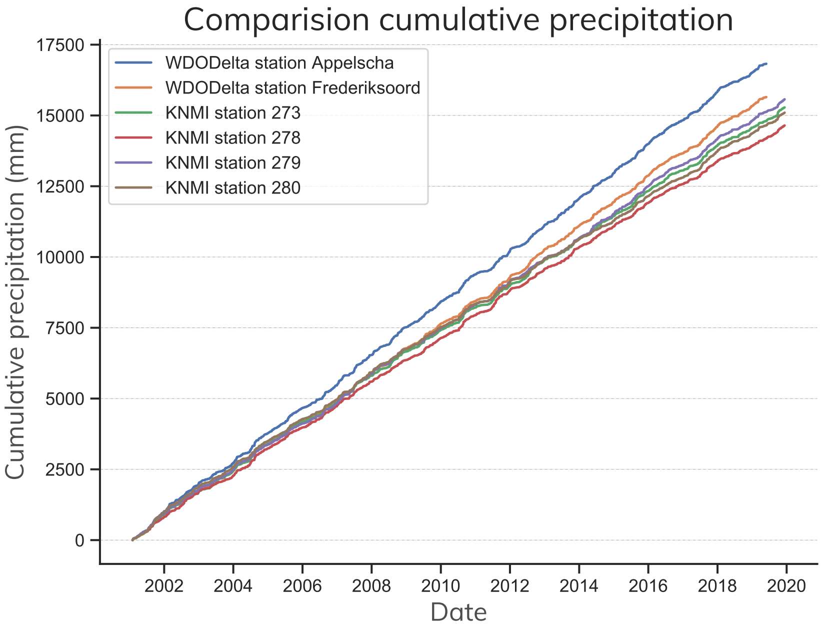


Figure 3.4: Cumulative precipitation of different rain gauge stations since February 2001.

of WDODelta, the null hypothesis H_0: a value from the WDODelta station is taken from the Normal distribution of the KNMI data, is tested. Or in other words: the station of WDODelta measures the same rainfall intensities as KNMI over longer periods. The cumulative values from February 1 2001 to June 6 2018 are used and the t-values where 13.88 for the station at Appelscha and 5.59 for the station at Frederiksoord. With $n = 4$ and $\alpha = 0.01$, $t_{\alpha,n} = 3.747$, so H_0 can be rejected. Therefore, the KNMI data is used instead of the WDODelta data.

As an additional test, the precipitation data has also been used to create a water balance in the Budyko framework. Both sets of precipitation data give plausible results (see Figure 3.7), although the position of the one with KNMI data is more likely.

The fluxes have been averaged per month of the year to create a water balance over an average year, which can be found in Figure 3.5. In general, in months with a higher potential evaporation than precipitation, the groundwater is depleted, whereas in the other months, the groundwater is resupplied.

It is possible that in the months with groundwater depletion, the water board is using artificial inflow to compensate for the water shortage. If that is the case, a higher discharge than expected, is possible during these months. Moreover, in the same months, many fields are sprinkled such that the actual evaporation is higher than could be expected and the open water levels are depleted, if no measures are taken.

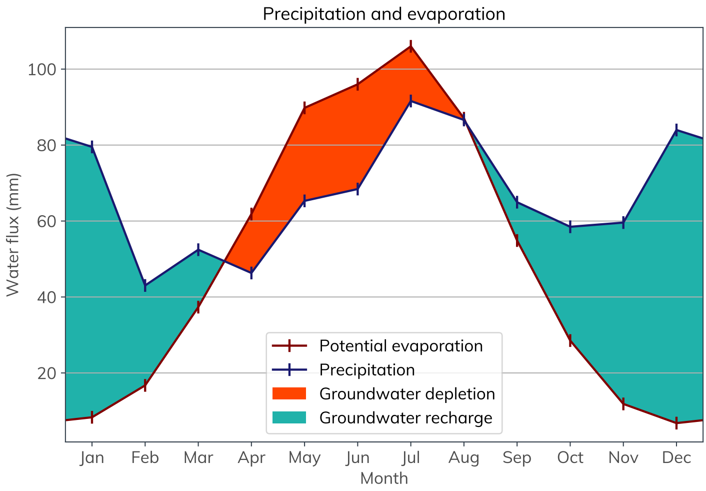


Figure 3.5: Precipitation and evaporation fluxes over an average the year

3.3.5. GROUNDWATER LEVEL

There are 1000s of groundwater measurement stations in the catchment, from both WDO-Delta, as well as the website DINOloket[6] from the Dutch scientific organisation TNO. These are groundwater wells from a wide range of sources. However, the data from these groundwater observation wells should meet three requirements, for it to be usable in a BN:

1. High temporal frequency (at least one measurement per subset period, see Section 3.3.2, but preferably over 100 per subset period)
2. Spanning the temporal range of the other parameters
3. Small number of errors

All, except for one, stations failed to meet these three requirements. Most had highly irregular data, or would only be available for a short period of time, and many had highly unlikely values. The one station that did meet all of the requirements was a water pressure station from WDODelta and can be seen in Figure 3.2. The fact that only one well met the requirements is highly unfavourable for the model, as water levels, and changes in water level, can differ heavily spatially. However, combining insufficient records has as an implication that different forces work on the model for different timesteps, which interferes with the IID assumption that is needed for a BN. In the water balance, however, many imbalances may arise from the fact that only one station is used.

As the goal for the water balance is to see water differences, the pressure should be translated to a water level, which can be done as follows:

$$h_w = \frac{p_w - p_a}{g}, \tag{3.2}$$

[6] https://www.dinoloket.nl/

where h_w is the water level, p_w is the water pressure and p_a is the atmospheric pressure. The latter is taken from KNMI station 279 as this was the only station in the close proximity that measures this data. Now, only an effective porosity factor is needed, which is chosen as 0.18, because it is an area with a fairly large amount of sand and hummus rich soils (see Section 3.2.5). The range of the effective porosity number is quite wide. The number 0.18 is arbitrarily chosen as one of the median numbers.

As BNs are able to use any number directly, the water pressure p_w is used directly in the BN model.

3.3.6. Surface water level

Surface water levels are water levels in rivers, canals, lakes etc. Levels measured near discharge measurement stations are highly correlated with the discharge. In a sense, they not so much predict discharges, but merely provide an early measurement. As the goal of this research is to make a predictive model, these kinds of surface water levels are not used in the model.

However, there is also a more managed region in the catchment. This is located in the (north-)western part of the catchment (see Figure 3.2). The surface water in this region is managed based on the desired groundwater level in that area. The management of the water level with weirs and some small pumping stations, potentially has a large influence on the discharge from that area as well. This region has one water level measurement station that had abundant frequent and longevity of recordings. This measurement location can be used in the BN, because BNs often also handle measurements that do not cover the whole area. This means that the water level is not representative for the whole catchment, but still adds an additional partial correlation - and therefore predictive power - to the model.

As this water level showed a lot of seasonality, it was regarded to implement a moving average filter. However, this was not implemented as deviations from this moving average can imply that there was an extreme event during that time. Only clear outliers that showed higher and lower levels than the measurement device can measure, have been removed.

3.3.7. Soil moisture

Soil moisture is an important hydrologic parameter, as it tells a lot about recent rainfall, evaporation and crop suction. In the catchment, no terrestrial measured, frequently sampled soil moisture datasets are available, that are representative of a large area of the catchment. Therefore, satellite data is used. As these satellites only measure the top part of the soil, the data mainly consists of water contents in the unsaturated zone. Due to its high frequent fluctuations, but limited volume, the soil moisture content is neglectable on a monthly timescale compared to other water fluxes and storage, and is therefore not included in the water balance (Section 3.4). However, soil moisture measurements can potentially give an indication of the discharge downstream. Therefore, soil moisture measurements are collected to be used in the BN.

There are a range of soil moisture satellite data products, but there is only one useful set that goes back to 2009: the Soil Moisture and Ocean Salinity (SMOS) satellite of by the European Space Agency (ESA) (European Space Agency, 2020). SMOS measures L-band brightness temperatures with a radiometer, which it uses to derive an estimate for the Soil Moisture (European Space Agency, 2017).

SMOS uses an Icosahedral Snyder Equal Area Earth grid, which consists of imperfect hexagons, with equally spaced cell centres at approximately 15 km distance (González-Zamora et al., 2015). Just two pixels cover more than 95 percent of the catchment area. These hexagons are irregular in nature and a detailed decomposition of this shape in relationship to the catchment shape is not assumed to deliver significant differences compared to just taking the average of these two points. Therefore, the latter approach is used in this research.

Due to its sun-synchronous orbit (European Space Agency, 2017), SMOS has an irregular temporal distribution, with sometimes multiple values per day, but also gaps of up to 10 days. This can cause problems when subsetting shorter time periods and therefore demands temporal averaging. For the first model made in this research, all the subsets did contain soil moisture data, but some were more frequent than others. Using a more frequent soil moisture product would be an important improvement to the model in the future. For now, adding the soil moisture only gives an indication of whether this could be useful in a model, and its contribution may not always be useful for each timestep.

3.3.8. NORMALISED DIFFERENCE VEGETATION INDEX (NDVI)

The NDVI is an indicative value for the so-called "greenness" of the vegetation in that area, with theoretical values between -1 and 1. It is defined as follows:

$$\mathrm{NDVI} = \frac{NIR - VIS}{NIR + VIS}, \tag{3.3}$$

where NIR stands for near infrared light and VIS for visible light. Chlorophyll, the pigment in plant leaves greatly absorbs visible light but the cell structure on the leaves strongly reflects near infrared light (NASA, 2000), resulting in a high NDVI when there is a lot of leaf activity. That is why NDVI is a great indicator of the condition of the fields and plant's need and availability of water. However, in common hydrologic models, the NDVI is not of use as it does not represent a water flux or storage. The BN, however, can deal with these kind of parameters. On the otherhand, the parameter is not nondescript to the user of the model, as it shows something in the network directly, in contrast to, e.g. only using a single band of radiation. Therefore, this has the potential to be an excellent parameter for a hydrologic model based on a BN.

Parameter	Band	Wavelength (nm)
NIR	Band 2	841876
NIR	Band 1	620670

Table 3.1: Wavelenghts used to construct NDVI in MODerate resolution Imaging Spectroradiometer (MODIS) (Didan et al., 2015; Earth Observing System, 2013)

In this research, MODIS from the National Aeronautics and Space Administration (NASA) is used as the source of NDVI, as this has a high density, assumed accuracy and is collecting data since 2000. MODIS' product MOD13A1 (Didan, 2015) is chosen, which has 16-day averages on a square grid of 500m (Didan et al., 2015). This is abundantly dense, as there are hundreds of pixels in the catchment, and the time average merely provides practical support as temporal subsetting becomes easier. Because NASA already averaged the raw data, some information has been lost here.

MODIS also has an advanced vegetation product, the so-called enhanced vegetation index (EVI). This has as an advantage that NDVI extremes are better projected. In this research, this is assumed to be futile because no extremes are expected in the case study area.

3.3.9. PROCESSING AND FILTERING

The data sources have been combined in Python, to be used for the BN and the water balance. To remove errors, the following operations have been performed:

1. The data has been summed/averaged as much as possible to remove non-bias errors. For the water balance, monthly averages were taken. As for the BN: for the discharge this meant that the maximum runoff was based on the MMDAD. The other variables were subset and averaged on a period before this event. In the initial model, this was 8 days plus the first half of the day of the MMDAD event. This number of days has been optimised in Section 7.3.

2. From the discharge data, values that had a larger negative discharge than -5 m^3/s were removed, as these did not fit in the distribution that followed from the other parameters.

3. Impossible values (often values that actually represent a NaN value) have been removed, such as negative precipitation depth (KNMI records precipitation below the detection limit as negative values or very low positive values).

4. Data has been spatially combined, as mentioned before.

It was out of the scope of this thesis to manually check for biases, so they were not removed. This does not matter for the BN, but it does matter for the water balance. Any other potential errors were also not removed, as it was not as clear as with the other cases that the observed effects were indeed due to errors.

For usage, all the parameters were converted to mm to use in the water balance. Where water fluxes were concerned, this meant that the catchment area was used to calculate the mm values. The precipitation was also changed to mm in the BN, because that is the widely used value for precipitation measurements.

3.4. WATER BALANCE

Creating a water balance is a useful method to verify the data, as well as to give insight into the climatology of the catchment. For the water balance, the fluxes that will be implemented in the BN will be used, supplemented with data that closes the balance as much as possible. For variables that will be included in the BN, such as the groundwater level, it can be considered to adding additional stations to create a more spatially frequent dataset, which can make for a more complete, less erroneous water balance. This has been disregarded because the focus of the water balance is to test the data and not to make a perfect balance as possible. See Table 3.2 for all of the fluxes. Other water fluxes have been neglected, such as seepage and infiltration (as these are difficult to measure), and data sources such as the NDVI or solar radiation that are no water flux are not included.

As the actual evaporation is not known, the potential evaporation is used initially.

Flux	In or out	Source
Discharge	Out	WDODelta
Precipitation	In	KNMI
Potential evaporation	Out	KNMI
Groundwater level difference	Out	WDODelta
Inlet Dieverbrug	In	WDODelta
Pumping station Bosweg	In	WDODelta
Groundwater extraction well Terwisscha	Out	AdviesCommissie Schade Grondwater

Table 3.2: Fluxes and storage in the water balance.

3.4.1. Monthly scale

By averaging all of the fluxes per month, and by neglecting soil moisture differences, a clear overview of the fluxes can be obtained for an entire year. Moreover, in a truly freely discharging system, the difference between incoming and outgoing fluxes should be approximately zero for each month. Large deviations from zero suggest measurement errors or incomplete information about the catchment. Moreover, as the actual evaporation is not known, the potential evaporation is used, which gives a outflow out of the catchment that can be too high.

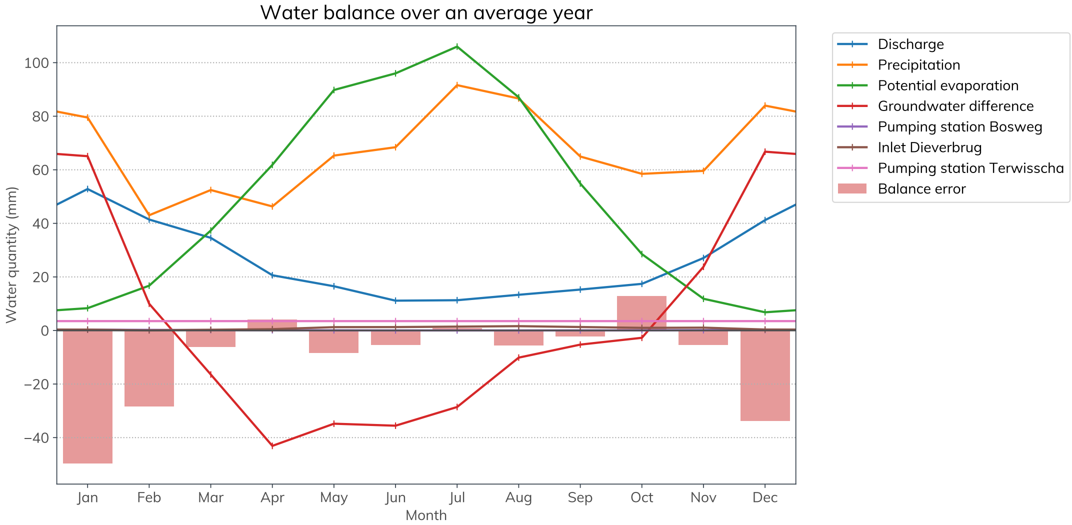


Figure 3.6: Water balance in the Vledder, Wapserveense and Steenwijker Aa over an average year.

In Figure 3.6, the average yearly water balance is shown. Upon consideration of the scale of this catchment, the fluxes of the pumping stations and inlets are deemed to be insignificant. Moreover, the balance between the fluxes is reasonably well recorded in the biggest part of the year, except for the winter months December to February. This implies that there is more water added to the storage, or discharged during the winter, than can be expected from the other fluxes. The reason that this happens is unclear.

3.4.2. Complete timeframe - Budyko framework

In hydrology, two indices have been constructed to get a sense of the climatology and hydrology of a catchment. The aridity index is constructed as the ratio between the potential evaporation and the precipitation in a catchment (E_p/P). It gives an indication of how dry the region is in which the catchment is situated. The evaporation index is the ratio between the actual evaporation and the precipitation (E_a/P). This gives a sense of how much of the water that comes into the catchment, leaves the catchment by means of evaporation. For a natural catchment, there are two limits for both of these indices. As the influence of the artificial influxes is neglectable (see Figure 3.6), the assumption is that these limits should hold true:

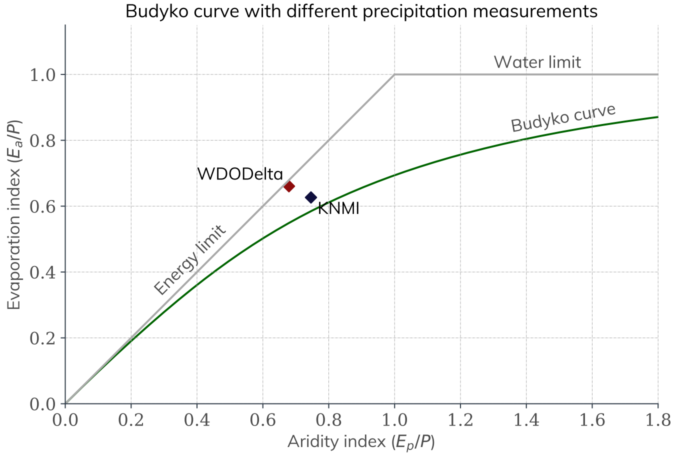


Figure 3.7: Catchment plotted for the aridity index versus the evaporation index, with different precipitation measurement sources. It can be seen that the WDODelta data plots just underneath the energy limit, which is possible but not common. Additionaly, the Budyko curve of Equation (3.5) is also plotted.

1. The actual evaporation cannot be higher than the potential evaporation, because no more energy is available. The limit is $E_a \leq E_p$, which gives the limit $E_a/P \leq E_p/P$. This is the so-called 'energy limit'.

2. The actual evaporation cannot be higher than the amount of precipitation, because not more water is available. The limit is then $E_a \leq P$, which gives the limit $E_a/P \leq 1$.

The Budyko framework is created for the whole time period of the data, rounded down on whole years. The difference in storage over this range can be neglected. The actual evaporation is calculated with the water balance, from the difference in the fluxes of Table 3.2, leaving the potential evaporation out of the equation and neglecting storage differences:

$$E_a = P - Q + \sum \mathbb{F}_{add,i}, \tag{3.4}$$

where Q is the discharge and $\mathbb{F}_{add,i}$ are the additional fluxes, with their appropriate sign. In Figure 3.7, the aridity index of the catchment has been plotted against the evaporation index, for two different precipitation measurements: the KNMI and WDODelta (see Section 3.3.4). Both measurements meet the energy limit, so cannot be rejected. Therefore, continuing to use the KNMI data because this would be more accurate, is still a assumption.

When many catchments are plotted in the same graph as Figure 3.7, these catchments seem to loosely follow a curve. At least six curves have been made to describe this curve (Arora, 2002). These are called Budyko curves. One example of such a curve, is the one proposed by Budyko and Miller (1974), which is the geometric mean between two other Budyko curves:

$$\frac{E_a}{P} = \left[\frac{E_p}{P}\tanh(\frac{P}{E_p})(1-e^{-\frac{E_p}{P}})\right]^{1/2}. \tag{3.5}$$

This function has also been plotted in Figure 3.7. It is clear that the KNMI precipitation dataset gives as a result that the catchment is a lot more average for its climatology, than position on the Budyko framework when the WDODelta precipitation dataset is used. This supports the preference for using the KNMI data over the WDODelta data to a small degree. A possible reason why the catchment has a higher actual evaporation than is expected form the Budyko curve, may be due to the high amount of agricultural fields and sprinkling of the fields. This should be visible in open water and groundwater levels.

4

INITIAL MODEL AND PERFORMANCE

4.1. INITIAL MODEL

In this research, many model parameters are analysed, to find the optimal model for this catchment and test whether copulas used are in agreement with the data. A first Bayesian network (BN) layout has been constructed to be used this optimisation. The data shown in Table 4.1 has been implemented in the first model, which looks like Figure 4.1. A correlation diagram can be found in Figure D.7 in the appendix. After this optimisation, in Chapter 8, a number of steps will be made to optimise the BN layout[1].

Variable	**Source**	**Frequency**	**Usage**	**Unit**
MMDAD	WDODelta	1/15 min.	Average	m^3/s
Precipitation	KNMI	1/hour	Sum	mm
Temperature	KNMI	1/hour	Average	C
Solar radiation	KNMI	1/hour	Average	$J/cm^3/h$
Soil moisture	SMOS (ESA)	irregular	Average	m^3/m^3
NDVI	MODIS (NASA)	irregular	Average	-
Groundwater levels	WDODelta	irregular	Average	kPa
Surface water levels	WDODelta	irregular	Average	m

Table 4.1: Variables used in BN models

Finally, here are some other, arbitrarily chosen parameters of the first model, which are tested in Sections 7.1 to 7.3:

- When sampling a conditioned copula, the expected value is taken as the *mean* of the sampled values, instead of the *median*

[1] Another order of methods is to first perform the initial steps of Chapter 8, which has some advantages in comparison to this order. However, this order is the one that has been used throughout the process of making the thesis. The model proposed in this section does not fully comply with these criteria, albeit not by much. It is assumed, however, that for most of the parameters, no significant changes in the results of the optimisation can be expected because of this discrepancy, as the initial and the final BN layout overlap considerably.

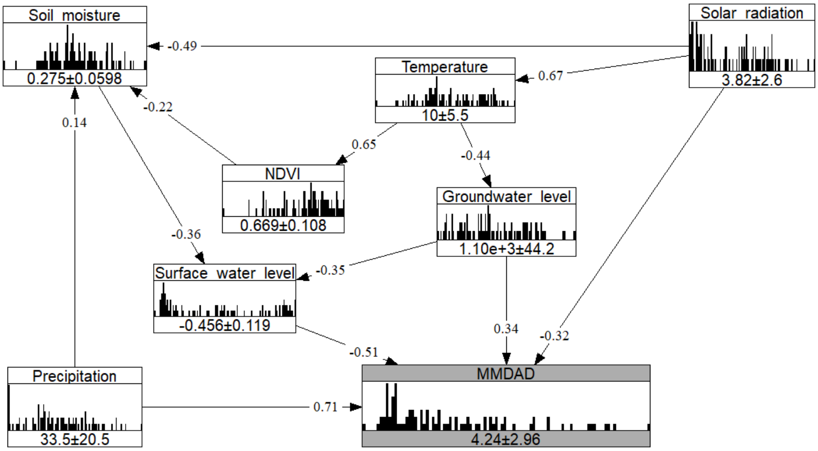


Figure 4.1: Initial BN model *it-0* that is used in this research, in Uninet. The numbers on the arrrows denote the correlation coefficients. See Appendix A for an explanation of the BN layout.

- The copulas are sampled 5000 times
- The other parameters use data from the 8 days before the highest discharge date

4.2. DETERMINING PERFORMANCE

In order to test whether parameters increase, or decrease the model performance, the definition of model performance needs to be determined. As the target variable of this research is the monthly maximum daily average discharge (MMDAD), the performance indicator evaluates the error between the prediction and the observations, in most of the cases in this research.

Two similar performance coefficients are discussed, to be used to connect a number to the performance of the model. This ensures that a great number of parameters can easily be fitted on the basis of the model performance, without ambiguity about what it means to have a good performing model.

4.2.1. NASH-SUTCLIFFE EFFICIENCY (NSE)

In hydrology, the NSE, is usually used to determine the performance of the model. The NSE is defined as follows (Nash and Sutcliffe, 1970):

$$NSE = 1 - \frac{\sum_{t=1}^{t_e} \left(Q_{sim}^t - Q_{obs}^t\right)^2}{\sum_{t=1}^{t_e} \left(Q_{obs}^t - \overline{Q}_{obs}\right)^2}, \tag{4.1}$$

where Q_{obs} is the observed discharge, Q_{sim} is the simulated discharge, the overlined $\overline{Q}_{obs}$ is the mean of the observed values and t_e is the final timestep of the data. The NSE has an upper limit of 1 (which is a perfect fit) and no lower limit. It highly depends on the purpose of the model and the quality of the data, but in general an NSE greater than 0.9 is regarded as a very good model, greater than 0.8 as a good model, and greater than 0.7 as a decent model.

4.2.2. KLING-GUPTA EFFICIENCY

An improved version of the NSE has been constructed by Gupta et al. (2009). This method contains separate elements for the difference in correlation (ρ_p), a measure to check the difference in volatility ($\boldsymbol{\alpha}$), and a measure to check the bias (β). The aggregated Kling-Gupta efficiency (KGE) is a useful way to get a quick overview of the model performance, but the separate building blocks give a more holistic view of the model performance and give room for a more purpose-dependent validation (Knoben et al., 2019). Therefore, this is the preferred test method in this research. A KGE score cannot be one-to-one compared with an NSE score (Knoben et al., 2019), however, the score limits are equal and the performance ranges are similar. The KGE is defined as follows:

$$KGE = 1 - \sqrt{(\rho_\text{p} - 1)^2 + (\boldsymbol{\alpha} - 1)^2 + (\beta - 1)^2}, \tag{4.2}$$

with ρ_p is the Pearson correlation coefficient (Equation (2.10)) and

$$\boldsymbol{\alpha} = \frac{\sigma_{sim}}{\sigma_{obs}}, \tag{4.3a}$$

$$\beta = \frac{\mu_\text{sim}}{\mu_\text{obs}}, \tag{4.3b}$$

where σ are the standard deviations of the simulations and observations, and μ is the mean (Knoben et al., 2019). In the case of validating discharges, μ_{obs} of Equation (4.3b) is equal to $\overline{Q}_{obs}$ of Equation (4.1).

4.3. k-FOLD CROSS-VALIDATION

In order to prevent overfitting, and similarly, to test whether the method can predict with data that it was not trained with, a k-fold cross validation is used in almost all of the tests.

In the method proposed in Section 2.6, there are two functions that have to be fitted:

1. The cumulative distribution function (CDF) fit function $F(v)$ for each $v_1, \ldots, v_n$ as defined in Equation (2.3).

2. The correlation matrix R, as defined in Section 2.6.2.

The CDF can easily be fitted differently per fold. However, the correlation matrix is fitted via the recursive Equation (2.17), which is done in Uninet. Constructing a new correlation matrix requires a whole new model in Uninet. Because of this, it is too time-consuming to create a new correlation matrix per fold. Therefore, the k-fold cross-validation is done only partially: on step 1. In Section 9.1, there is a single k-fold test for both the fitted items.

Another method to do cross validation could be Monte Carlo cross validation. This method consists of randomly selecting a test and a training set, running the test and repeating this process n number of times. However, as the correlation matrix R is determined from all of the data in the range, it is more consistent to take all of the timesteps into account. This is not certain with the Monte Carlos method. Moreover, not all the timesteps are weighted equal in this test. Therefore, initially the k-fold cross validation is used in this research.

In many cases, just a single k-fold test does not suffice, because having random folds gives less certainty over the median performance. That is why, in many cases in this research, a 5-fold test is combined with a 10 or 20 times Monte Carlo repetition, in order to get repeatable results.

4.4. PERFORMANCE PER OBSERVATION

In several cases, it is useful to not only look at a single error number, for example when a clear differentiation can be made in predictions of high observations and low observations. In Section 6.6, the error is also plotted against the corresponding observation. This gives an overall image of the performance of the model over the observations.

5

TESTING THE COPULA ASSUMPTION

5.1. INTRODUCTION

In Chapter 2, it is determined that for this case study, the Gaussian copula modelled through the multivariate normal (MVN) method, it the likely the most useful method implement copulas in the Bayesian network (BN). The assumption that the data of the case is usable in a BN and that the Gaussian copula suits the joined probability distributions of variable pairs, is tested in this chapter. The following tests are performed:

(i) Autocorrelation test: The purpose of the BN is to model the complete joined dependence structure of the variables. This means that the parent variables that are no child should behave like independent variables. The marginal distributions of these variables should be independent and identically distributed (IID), as these are all uniformly distributed. The notion of independence also holds that the variables should be independent in time. This assumption is flawed, as the variables show a degree of seasonal influence, such as for example the solar radiation. Moreover, for the child variables, it means that they should be independent given its parents Paprotny and Morales-Nápoles (2017).

As the target of the variable selection in Section 3.3 was to select all significant influences on the catchment, the only other dependence should come from the other variables. If there is no influence from other parameters than those that are used in the BN or from itself, the variable is regarded as conditionally independent. This assumption is challenged because the values come from time series, which often show serial correlation. This is already counteracted to a degree by taking the monthly maximum and subsetting the data, see Figure 3.3. How much of the self-dependence still remains is tested with an autocorrelation test in Section 5.2. Moreover, for the monthly maximum daily average discharge (MMDAD), it is also tested how much autocorrelation remains after the effect of its parent variables is removed, through a partial autocorrelation test.

Testing whether other variables, of which data is known, also have an influence a certain variable is tested by calculating the correlation coefficients between these parameters. For example, air pressure had an absolute correlation of less than 0.1 with the discharge, and therefore it is regarded to not influence the model. Other parameters, such as a discharge close upstream of the measurement station, were regarded as having too much predictive power and therefore being more an additional measurement than a useful variable for the BN. An exhaustive analysis about this theme is outside the scope of thesis.

(ii) Test fit of copula and show indications of tail dependence: Secondly, in this research, the Gaussian copula is used. It is tested whether this is one of the copulas that fit the data best in Section 5.3 and which positions the copula over- and underestimates the correlation in Section 5.4. Lastly, it is tested whether there is tail dependence in the copulas, since the employed Gaussian copulas do not model this. If there is tail dependence, which type of copula would have modelled this better, is explored. This is featured in Section 5.5.

Testing copulas requires the usage of uniform marginals that can be deducted from the variables with a cumulative distribution function (CDF) (see Equation (2.3)). In Chapter 6 the best probability distribution is selected. In this chapter, the empirical cumulative distribution function (ECDF) is used (see Equation (6.1)). All the tests are conducted on pair-copulas.

5.2. Autocorrelation test

5.2.1. Unconditional autocorrelation

The autocorrelation $\rho_{auto,l}$ of a variable is the correlation between a value and the value that happened l timesteps before that, for all values. These l timesteps are called the *lag*.

$$\rho_{\text{auto}}(l) = \frac{1}{n-l} \sum_{i=1+l}^{n} \rho_{\text{p}}(v_i, v_{i-l}), \tag{5.1}$$

where n is the number of values of a variable and $\rho_{\text{p}}(x, y)$ is Pearson's correlation coefficient between two variables x and y (see Equation (2.10)). In the Figure 5.1 the autocorrelation test is done for all variables that are used in the initial model. Three lags have been plotted. A lag of 1 month usually yields the highest autocorrelation coefficient and this lag shows a lot about the general volatility of the variable. Lags of 6 and 12 months denote the seasonal influence of parameters.

Concluding from Figure 5.1, the variables Soil Moisture, MMDAD and especially the precipitation have a low autocorrelation and can be used as virtually temporally independent values. However, most of the other parameters show a large seasonal effect. This means there is less information that can be extrapolated from the dataset than without a seasonal effect. This potentially makes the correlation between the variables stronger then its actual influence is. However, it can also be argued that the season is well taken into account when conditioning, and because the main forcing factors on the model of the season (temperature, solar radiation, water levels) are all implemented in the BN. The groundwater level has a slightly lower seasonal effect and can therefore be regarded as being approximately unconditionally independent in time.

5.2.2. Partial Autocorrelation

The assumption of the BN is that the data is conditional independent to the previous values, given its parents (Paprotny and Morales-Nápoles, 2017). Therefore, it is also important to look at autocorrelation of a variable with the effect of its parents removed. For the target variable of the MMDAD, the partial Pearson's correlation has been calculated with Equation (2.17), to show this effect. Pearson's correlation coefficient is used, instead of for example Spearman's

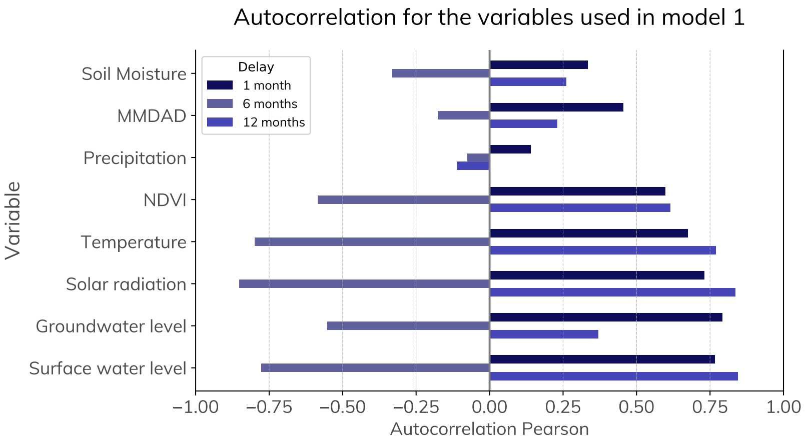


Figure 5.1: Autocorrelation for the variables used as processed in model 1, for the delays of 1 month, usually the highest autocorrelation and an important factor in showing volatility, and for 6 and 12 months, to see seasonality effects.

correlation, because this makes it more comparable to the previous question and is more widely used. The conditional variables were the precipitation, surface water level, groundwater level and solar radiation, as can be seen in Figure 4.1. The results of the test can be found in Table 5.1. This table shows that the autocorrelation for the long periods of half a year and a whole year have become very low and these are effectively independent to the measurement with lag 0. This means that the seasonality has been carried by the other variables. Moreover, the partial correlation for the a single month has been marginalised and only plays a small role. Therefore, it can be concluded that the way the MMDAD variable is set up, makes it effectively conditionally independent.

Lag (months)	Pearson's corr. coeff. (ρ_p)
1	0.204
6	0.066
12	0.092

Table 5.1: Partial autocorrelation for the MMDAD.

5.3. MULTIDIMENSIONAL CRAMÉR-VON MISES TEST

5.3.1. THEORY

To check if a copula fits well to the data the goodness of fit test proposed by Wang and Wells (2000) is implemented. They propose a multidimensional Cramér-von Mises statistic $\mathfrak{S}_n$:

$$\mathfrak{S}_n = n \int_1^n \left(C_{emp}(u) - C(u)\right)^2 \mathrm{d}C(u), \tag{5.2}$$

where n is the number of samples, $C(u)$ is the CDF of the d-dimensional copula to test against and $u \in [0,1]^d$ to call the copula. Lastly, $C_{emp}(u)$ is the empirical CDF copula and is defined as follows (Genest and Remillard, 2008):

$$C_{emp}(u) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left(\frac{\mathbb{R}_i}{n+1} \leq u\right), \tag{5.3}$$

where $\mathbb{R}$ is the rank vector, with size d, of each data point and $\mathbf{1}(\cdot)$ is the indicator function.

Genest and Remillard (2008) propose a solution for Equation (5.2):

$$\mathfrak{S}_n = \sum_{i=1}^{n} \left(C_{emp}(u) - C(u)\right)^2. \tag{5.4}$$

In this report, an alteration of the Cramér-von Mises statistic is presented, which is based on the root-mean-squared deviation (RMSD):

$$\mathfrak{S}_{RMSD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(C_{emp}(u) - C(u)\right)^2}. \tag{5.5}$$

It is important to note that it is a mean over the data points, not over the XY-plane of the copula. It is, in a sense, scaled, by the chances of occurring from the data: more differences are taken into account where there are more data points.

Furthermore, (Genest and Remillard, 2008) also proposes a bootstrapped method to test H_0, that the combined data follows a Gaussian copula. In this research, the one-level bootstrap method posed in this paper, is used with Equation (5.5) instead of Equation (5.4). The one-level method is sufficient, as it performs equally well as the two-level method. This method is defined as follows:

1. Create an empirical copula $C_{emp}(u)$ with the data.
2. Calculate $\mathfrak{S}_{RMSD}$ via Equation (5.5).
3. Pick a large N and repeat the following steps for every $i \in \{1,2,\dots,N\}$.
   (a) Generate a random sample $u^*_{1,i},\dots,u^*_{n,i}$ $(u^* \in [0,1]^d)$ from the copula $C(u)$
   (b) Calculate the empirical copula of this random sample $C^*_{emp,i}(u)$ via Equation (5.3).
   (c) Generate a copula $C^*_i(u)$ with the same method as the original copula.
   (d) Compute $\mathfrak{S}^*_{RMSD,i}$ via Equation (5.5), with $C^*_{emp,i}(u)$ and $C^*_i(u)$ as respective $C_{emp}(u)$ and $C(u)$.
4. An approximation for the p value, the probability of finding at least $\mathfrak{S}_{RMSD}$, assuming that the H_0 is correct, is then defined as follows:

$$p \approx \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left(\mathfrak{S}^*_{RMSD,k} > \mathfrak{S}_{RMSD}\right). \tag{5.6}$$

Note that the results for p are equal when using Equation (5.4) in comparison to Equation (5.5).

5.3.2. RESULTS

The measure proposed in Equation (5.5) is calculated for all of the presented copulas in Section 2.3. For the marginals, the empirical CDF is used. In Chapter 6, two other methods are proposed for the marginal distributions. The average of the values of $\mathfrak{S}_{RMSD}$ per type of copula is presented in Figure 5.2. See Table F.1 in the appendix for the results per connection.

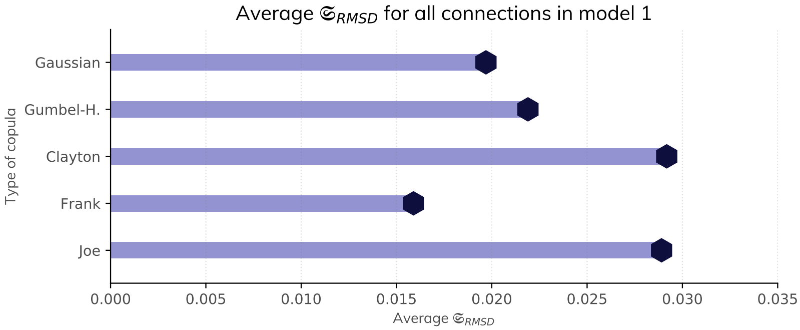


Figure 5.2: Results of the 2-dimensional CvM test for all connections in the BN of model 1. The $\mathfrak{S}_{RMSD}$ value for each variable combination and all copulas can be found in Table F.1 in the appendix.

It is clear that the Frank copula has the best fit in general, for this test, but a close second is the Gaussian copula. This is logical, as they have a similar configuration, see Figure 2.2 and Figure 2.6. Both of these copulas fit about twice as well as the other three copulas.

For all variable pairs, the p-value has been calculated with $N = 500$. It is decided that for H_0 that the empirical copula follows a similar joined distribution as the copula, $\alpha = 0.05$. The p-value has been calculated for all copulas described in Section 2.3. Sampling (step 3a) has been done in the Python module `pycopula` in a way described by Hofert (2008), for all types of copula except for the Joe copula. This has been done because results for $\mathfrak{S}_{RMSD}$ already showed bad fits, and on top of that, because sampling from the Joe copula is not defined in `pycopula`. In Figure 5.3 the number of rejected bivariate copulas is shown out of the 24 combinations that are in Model 1. It is clear that the most bivariate distributions could follow the Frank copula, whereas for the Gaussian and the Gumbel-Hougaard copula, many fits can be rejected. The Clayton copula seems to make hardly any plausible fit. The complete values can be found in Table F.2, in the appendix.

5.4. ABSOLUTE DIFFERENCES

A visual tool to determine the consequence of the suboptimal fit of the copula is to calculate

$$\Delta C = C_{emp}(u) - C(u), \tag{5.7}$$

for each of the values of u_1 and u_2 in the dataset. This is similar to the multidimensional Cramér-von Mises test (Section 5.3). An example plot between the empirical copula of the NDVI variable and the Solar radiation variable of model 1, and a fitted Gaussian copula, can be found in Figure 5.4.

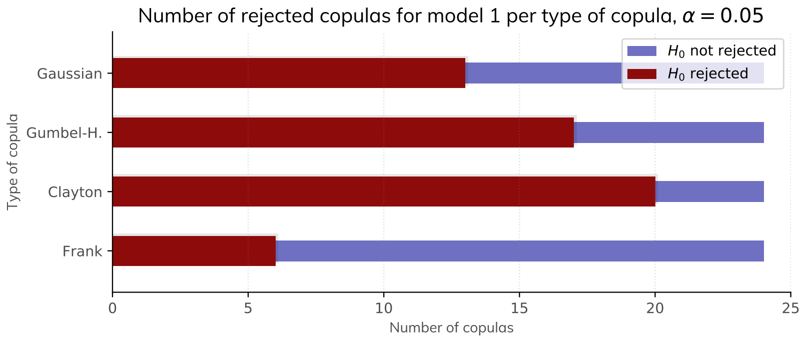


Figure 5.3: Number of bivariate copulas that can be rejected based upon the p-value method proposed by Genest and Remillard (2008) stated in this section, per type of copula tested on all combinations in model 1. The p-value for each of the variable combinations and all copulas can be found in Table F.2 in the appendix.

Figure 5.4 shows that, based on the data, in the north west and south east corners, the copula underestimates the chances of occurrence, whereas in the middle, the copula overestimates these chances. This tool is not conclusive in what copula to use, but based on the copula, it shows where errors might arise.

5.5. QUADRANT PEARSON CORRELATION

5.5.1. TAIL DEPENDENCE IN DATA

The assumption that the parameters follow the Gaussian copula, can be verified with another method: the investigation of its tail behaviour. Copulas show different type of tail behaviour, for example Archimedean copulas have different correlations close to one corner of the model in comparison to the opposite corner. Tail dependence can intuitively be defined as the probability that U_1 reaches extremely large values, given that random variable U_2 obtains extremely large values. The closer to a corner of the copula, the larger the chance of the other variable reaching an extreme value as well. The Gaussian copula does not have tail dependence (Paprotny, 2017). To test this for each combination of variables, the data is converted via the uniform values, to a standard normal distribution for both variables. Then, the data is divided into quadrants, with the dividers at the x- and y-axes. Next, the Pearsons correlation coefficient is calculated for all of these quadrants, as well as for the whole dataset. When the absolute correlation coefficient for one of these quadrants is higher than the total quadrants, a tail dependence can be expected. (Joe, 2015).

Of all the combinations tested, 16.7% of the quadrants had a higher correlation coefficient than the overall correlation. This suggest that many quadrants had no significant tail dependence. However, 11 of the 24 combinations had at least one quadrant that showed tail dependence. This implies that using the Gaussian copula for these combinations is imperfect. See Table F.3 in the appendix for the complete results.

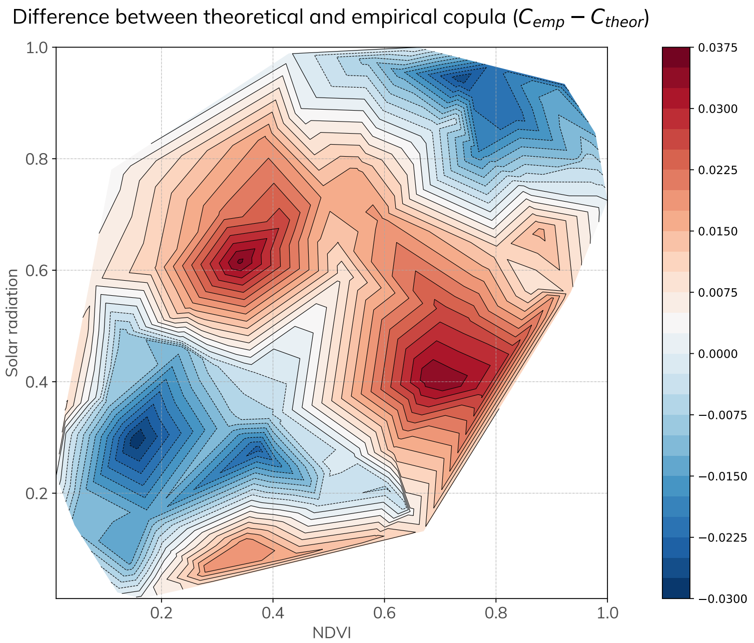


Figure 5.4: Absolute differences for the empirical copula of the variables NDVI and Solar radiation of model 1 in comparison with a fitted Gaussian copula.

5.5.2. TAIL DEPENDENCE FOR COPULA TYPES

Calculating the correlation coefficient of quadrants is also a way to see whether copulas fit the data well regarding tail dependence (Morales Nápoles, 2019). The test works by comparing the quadrant correlations for the dataset with quadrant correlations with samples from a fitted copula. It is performed as follows:

1. Calculate the quadrant correlations for the two variables.
2. Fit a copula of a certain family to the two variables.
3. Generate n samples from this copula
4. Use the same method to normalise the n sample variables
5. Calculate the difference in correlation coefficient per quadrant

In Figure 5.5 the average result per copula type can be found. It is clear that again, the Frank copula performs best, closely followed by the Gaussian copula. The Clayton copula performs better in this test than in the Cramér-von Mises test (see Section 5.3). The relatively lower correlation in the NW, NE and SE quadrants were similar to some of the data combinations. See Table F.4 in the appendix for the complete results.

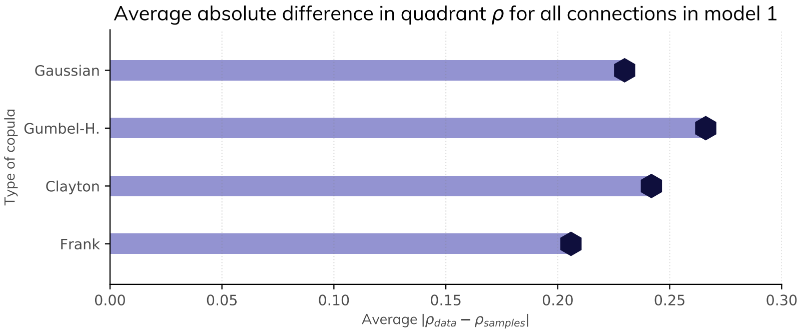


Figure 5.5: Average absolute difference in correlation coefficient per quadrant, per copula type.

6

Selection and implementation of the marginal distribution

The marginal distribution, used in the copula with Equation (2.3), can be implemented in several ways, empirically from the data or theoretically, with a fitted function. The advantage of fitting a function is that it can smoothen out the empirical function, when it does not have abundant data, and it has the possibility to predict values smaller or larger than ever recorded. Essential is that the function that is used to convert the variables to uniform marginals, is actually a probability distribution function, of which the cumulative distribution function (CDF) is used. Otherwise, it is possible that it will predict wrongly outside of the bounds of the previous data or that it's non-invertable, which makes that calculation the value from the respective uniform marginal has become ambiguous for some values.

A small exploratory study has been conducted in implementing conventional probability distributions, such as the beta, gamma, gumbel, rayleigh and normal distributions, as CDF for the marginal variables. This often gave poor fits and the prediction of the monthly maximum daily average discharge (MMDAD) became significantly worse than using the empirical CDF (see Section 6.1). Therefore, it was presumed that a more volatile CDF was needed for the marginals. As these volatile CDFs follow the empirical distribution very closely, it is also assumed that selecting different CDFs for the different variables does not make a large improvement to the model. Therefore, all variables are applied with the same probability distribution in each of the tests. Moreover, as the central target of this research is for the Bayesian network (BN) to predict the MMDAD well, all the testing is done a posteriori by optimising the Kling-Gupta efficiency (KGE) of predicting the MMDAD with the BN that is introduced in Section 4.1 for each of the parameters.

Three different methods are introduced to fit the marginals in this chapter. For each of these methods, its parameters are optimised. Afterwards, the implementation of the inverse of these function ($F^{-1}(\cdot)$) is discussed. In Section 6.5, the optimal method for this thesis is chosen.

6.1. Empirical cumulative density function

The marginal distribution can be implemented as a step function CDF:

$$F_{\text{emp}}(v) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left(\frac{v_i}{n+1} \leq v\right), \tag{6.1}$$

where n is the number of samples in the dataset, $v_1, \ldots, v_n$ are all the values in the dataset and $\mathbf{1}(x)$ is the indicator function. This is the one-dimensional case of Equation (5.3).

6.2. Altered logistic function

The logistic function is an S-shaped (sigmoid) function which functions as a cumulative density function. It is defined as follows:

$$F(v) = \frac{\alpha_0}{1+e^{-\alpha_1(v-\alpha_2)}}, \tag{6.2}$$

with α_0, α_1 and α_2 as constants. To be used as a CDF, α_0 should be 1. This leaves the function with only two more constants, which is too little to make a close fit to the data. Therefore, a polynomial is added to the function:

$$F_{\text{logi}}(v) = \frac{1}{1+e^{\alpha_1(\mathbb{P}(v)-\alpha_0)}}, \tag{6.3}$$

where $\mathbb{P}$ is a K-parameter ($2K+1$-degree) odd polynomial:

$$\mathbb{P}(v) = \sum_{i=0}^{K}\alpha_{i+3}\cdot(v-\alpha_2)^{2i+1}, \tag{6.4}$$

and $\alpha_0 \ldots \alpha_{K+3}$ are the constants. As $F_{\text{logi}}(v)$ represents a cumulative density function, it should be ever increasing. Therefore, $F'_{\text{logi}} = f_{\text{logi}}(v) \geq 0$ for $v \in (-\infty, +\infty)$, which gives[1]:

$$\begin{aligned} f_{\text{logi}}(v) &= \frac{d}{dv}F_{\text{logi}}(v) \\ &= \frac{-\alpha_1\cdot\mathbb{P}'(v)\cdot e^{\alpha_1(\mathbb{P}(v)-\alpha_0)}}{\left(e^{\alpha_1(\mathbb{P}(v)-\alpha_0)}+1\right)^2} \geq 0 \text{ for all } v \in (-\infty, +\infty). \end{aligned} \tag{6.5}$$

As $e^{\alpha_1(\mathbb{P}(v)-\alpha_0)} \geq 0$, now the condition is that $-\alpha_1\cdot\mathbb{P}'(v) \geq 0$. Since $\mathbb{P}(v)$ only consists of odd powers, $\mathbb{P}'(v)$ is made of even powers. Now, there are two solutions which make $-\alpha_1\cdot\mathbb{P}'(v) > 0$:

$$\alpha_1 \leq 0, \quad \alpha_3 \ldots \alpha_{n+3} \geq 0, \tag{6.6a}$$

$$\alpha_1 \geq 0, \quad \alpha_3 \ldots \alpha_{n+3} \leq 0, \tag{6.6b}$$

which are equal to each other because of the way Equation (6.3) is constructed. In this thesis, the condition of Equation (6.6b) is used. Furthermore, the formula should be asymptotic on 0 and 1, which is the case using the conditions of Equation (6.6):

$$\lim_{v\to+\infty} F_{\text{logi}} = \frac{1}{1+e^{-\infty}} = \frac{1}{1+0} = 1, \tag{6.7a}$$

[1] See Equation (C.4) in the appendix for the derivation

$$\lim_{v \to -\infty} F_{\text{logi}}(v) = \frac{1}{1+e^{+\infty}} = \frac{1}{1+\infty} = 0. \tag{6.7b}$$

Therefore, this can be used as a cumulative distribution function.

6.2.1. INITIAL VALUES FOR FITTING THE CDF

In order to get a stable curve fit algorithm as described in Section 6.4.1, which always finds an optimum, the initial values should be determined well. By testing configurations that made a very globally similar distribution for each of the variables, it was determined that the following parameters deliver a stable fit in almost all cases.

$$\begin{aligned}
\alpha_0 &= 0.2 \cdot \min(V)\frac{B}{2}, \\
\alpha_1 &= \frac{0.07}{B}, \\
\alpha_2 &= \min(V)\frac{B}{2}, \\
\alpha_3 &= -80 \cdot B^{-0.5}, \\
\alpha_4 &= -1 \cdot B^{-0.5}, \\
\alpha_5 &= -0.01 \cdot B^{-2},
\end{aligned} \tag{6.8}$$

where B is defined as $\max(V) - \min(V)$, and $\max(V), \min(V)$ respectively the maximum and minimum value of the variable V. Using 7 parameters (up to α_6) was not implemented as this delivered a power of 7 in the odd polynomial Equation (6.4), which proved to be highly unstable.

6.2.2. OPTIMAL NUMBER OF FIT PARAMETERS

The optimal number of parameters is determined a posteriori by calculating the KGE for each number of parameters with a k-fold cross validation, in contrast to determining the optimal number of parameters a priori via, for example, either AIC or BIC. This is the case because the central objective is to find the optimal BN, not the optimal $F(v)$ fit. It was outside the scope of this research to test whether a better performance could be acquired by using diffent numbers of parameters per variable.

For each number of parameters, the model was tested with 5 folds and the experiment was repeated 10 times. Therefore, in total 50 KGEs were calculated. A boxplot has been made of the KGE per fold in Figure 6.1. It is clear that at least five parameters are needed to acquire a good model performance. However, using six parameters does not add anything to the model performance. Therefore is it decided to use five parameters for the altered logistic function.

6.3. GAUSSIAN MIXTURE MODEL

The Gaussian mixture model is a combination of multiple Gaussian distributions. It has the ability to fit well to various empirical distributions. It also has an straightforward description of a probability density function (PDF) as well as a CDF. The PDF of the function for K normal

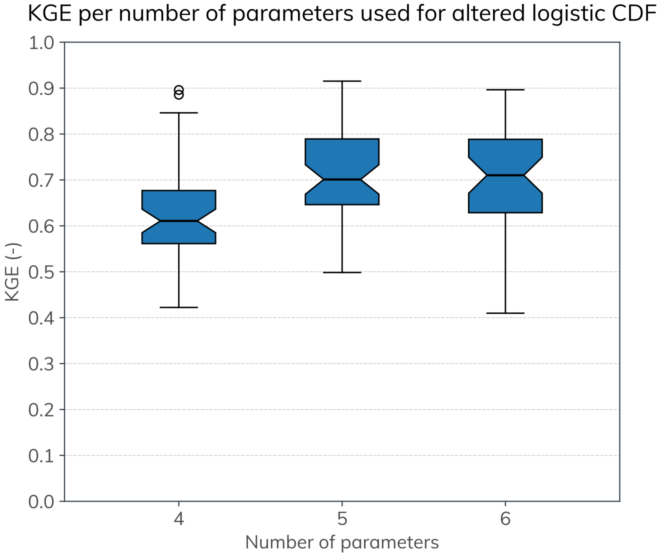


Figure 6.1: Boxplot of KGE per fold, per number of normal distributions for the Gaussian mixture model

distributions is defined as follows:

$$f_{\mathrm{gm}}(v) = \sum_{i}^{K} \frac{\alpha_i}{\sigma_i} \varphi\left(\frac{v-\mu_i}{\sigma_i}\right), \tag{6.9}$$

where $\varphi(\cdot)$ is the PDF of the standard normal distribution ($\mathcal{N}(0,1)$), μ_i the optimal mean, and σ_i is the optimal standard deviation of that gaussian. Now the CDF is determined as:

$$F_{\mathrm{gm}}(v) = \sum_{i}^{K} \frac{\alpha_i}{\sigma_i} \Phi\left(\frac{v-\mu_i}{\sigma_i}\right), \tag{6.10}$$

with $\Phi(\cdot)$ as the integral (CDF) of $\varphi(\cdot)$. The conditions are that:

$$\sum_{i}^{K} \alpha_i = 1, \tag{6.11a}$$

$$\alpha_1, \ldots, \alpha_K \geq 0. \tag{6.11b}$$

Now, because $\varphi(\cdot) > 0$ and Equation (6.11b), $f_{gm}(v) > 0$ for for all $v \in (-\infty, +\infty)$. Hence, the first condition of being a proper probability distribution is met. This also implies that $F_{gm}(v)$ is ever increasing. Moreover, as $\int_{-\infty}^{+\infty} \varphi(\cdot) = 1$, or symmetrically, $\lim_{v\to-\infty} \Phi(v) = 0$ and $\lim_{v\to+\infty} \Phi(v) = 1$, the total probability is 1, which satisfies the second condition to be a probability distribution.

As an additional advantage, in selected cases, the different normal distributions could be descriptive of an underlying process in the data. These cases consist mostly of m independent, underlying processes that follow a more or less normal distribution, which corresponds to the number of normal distributions that is selected.

6.3.1. INITIAL VALUES FOR FITTING THE CDF

Similarly to the altered logistic model, empirically, it was determined that the following initial values give stable fits in all cases:

$$\begin{aligned} \mu_i &= B\frac{i+1.1}{1.008K} + \min(V), \\ \sigma_i &= \frac{i+3}{0.15K \cdot B}, \\ \alpha_i &= \frac{i+5}{6}, \end{aligned} \tag{6.12}$$

where K is the number of Gaussians, B is $\max(V) - \min(V)$, and $\max(V), \min(V)$ respectively the maximum and minimum value of the variable. In the code, Equation (6.11a) is acquired by dividing between the sum of the αs.

6.3.2. OPTIMAL NUMBER OF NORMAL DISTRIBUTIONS

With a similar method as described in Section 6.2.1, the Gaussian mixture model was calculated for one to five normal distributions. As each normal distributions has three parameters (α_i, μ_i and σ_i), this was a total of 3 to 15 parameters.

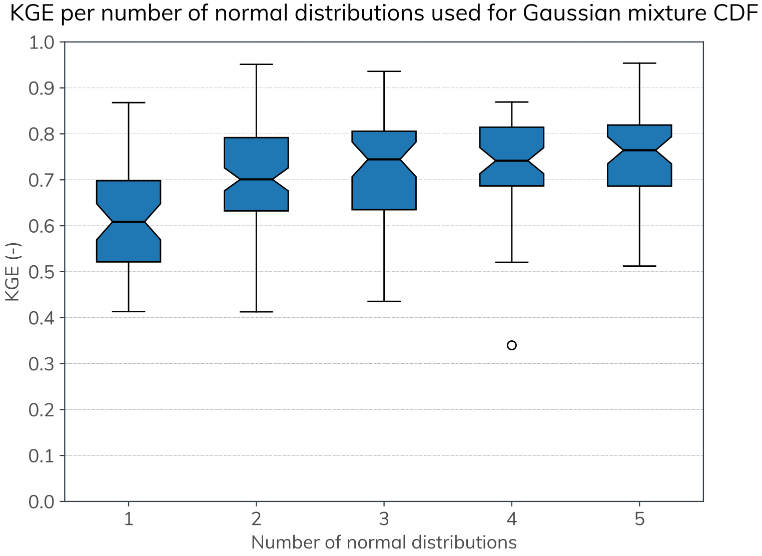


Figure 6.2: Boxplot of KGE per fold, per number of normal distributions for the Gaussian mixture model

In Figure 6.2 a boxplot is made of the average KGE per number of normal distributions. Upon observing this figure, using only one normal distribution is too low, but, using two does already perform really well. In this research, it is chosen to use three normal distributions, as the median is higher than with two distributions, and the performance seems to improve less noticeably from this number of parameters onwards.

6.4. INVERSE CUMULATIVE FUNCTION

For both theoretical distribution functions, the altered logistic function (Section 6.2) and the Gaussian mixture model (Section 6.3), the inverse function ($F^{-1}(\cdot)$) is not defined for all parameters α. Therefore, the model is first sampled regularly with 1,000,000 points over the x-axis and then linearly interpolated. To be able to extrapolate, the minimal value of the points over the x-axis is 20% times the difference between the lowest and the highest value, lower than the lowest measured value, and goes up to 20% times this range higher than the highest measured value. This interpolation has an especially good resolution in the margins, as the derivative here is close to 0. For the empirical cumulative distribution function (ECDF) (Section 6.1) the value with the uniform value closest to the sought after value is chosen.

6.4.1. FITTING THE CDFS TO THE DATA

In the case of unlimited parameters, fitting of the constants $\alpha_0 \ldots \alpha_{n+3}$ is done in Python with a trust region reflective (TRF) algorithm, because it is a bounded problem (Branch et al., 1999).

6.5. PREFERRED CDF

6.5.1. OPTIMAL FIT

A k-fold test with 5 folds has been performed for model 1 with the ECDF, the altered logistic CDF and the mixed Gaussian CDF, as fit function for all parameters. This process was repeated 20 times, such that there were 100 test results for each of the fit functions. The results can be found in the boxplot in Figure 6.3.

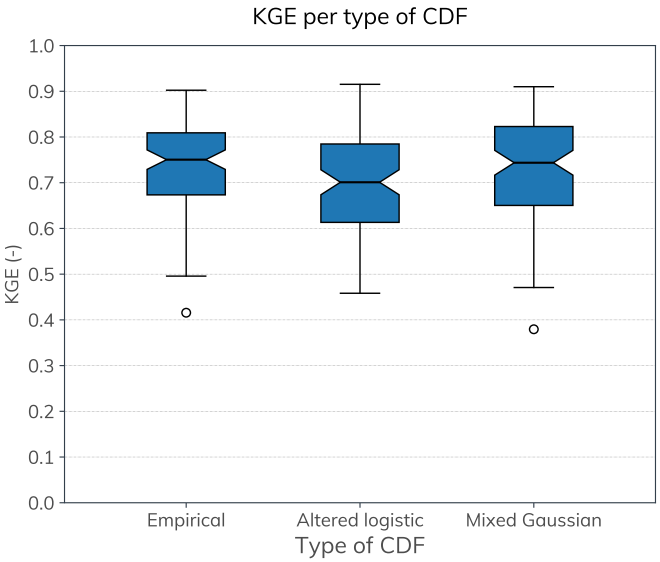


Figure 6.3: Boxplot of KGE per fold, per type of cumulative density function described in this chapter.

It can be seen from the graph that the ECDF and mixed Gaussian perform similarly, whereas the altered logistic function performs slightly worse. The range of the results from the mixed Gaussian CDF is a little wider than than from the empirical CDF, but not significantly.

STABILITY RESULTS

All probability distributions had some folds that performed poorly, but the number of folds that did was not significantly different between these distributions. However, the altered logistic CDF had three folds in which one of the variables could not be fitted to the function. This makes this distributions less stable.

6.5.2. EXTRAPOLATION

The functions that mimic the empirical distributions have the possibility to extrapolate data. Therefore, it is possible that they can calculate higher discharges than have ever been measured before. However, it is not known how well they do this.

6.5.3. COMPUTATIONAL DURATION

For all of the methods, the time to fit all of the CDFs and test one fifth of the model ($k = 5$ for all tests), was less than 10 seconds[2]. This is a lot shorter than for example running a SOBEK model (which can take from hours up to days, and fitting has to be done partly manually), and is therefore regarded as an insignificant amount of time. However, when larger models with more parameters than proposed in this thesis are used, the difference in time of fitting can be of importance. Performing a run as mentioned in this section took on average 3.5 seconds for the ECDF, 2.4 seconds for the altered logistic CDF, and 5 seconds for the mixed Gaussian CDF.

6.5.4. CONCLUSION

As the ECDF and the mixed Gaussian CDF performed similar when used to predict the MMDAD with the BN proposed in Section 4.1, but the mixed Gaussian has the potential to extrapolate, this is the one that is chosen for the rest of this research.

6.6. SHIFTING THE CDFS TO INCREASE EXTRAPOLATION

The ECDF's lowest value is exactly 0 and its highest value is exactly 1. This causes the fitted CDFs to fit very close to 0 and 1. Despite these values never actually touching the limits 0 and 1, and thus being able to extrapolate, this extrapolation is very rarely happening. The reason for this, is that the curve fit moves the function as close to 0 as possible on the lowest value of the ECDF and as close to 1 as possible for the highest value. Suppose that, in a fitted CDF, the uniform value of the maximum value, is 0.999. Then there is only 0.1% of the uniform space left for extrapolation on the high end. For predicting relatively high MMDADs, for example, it can be useful to be able to get more samples from extrapolations.

As a consequence, the method underestimates the high discharges (see the blue dots in Figure 6.7). In general, high discharges are most prone to cause floods. Therefore, underestimating these is unfavourable. It would be beneficial for these regions, therefore, for the method to extrapolate more.

[2] See Appendix G.2 for the hardware used for these calculations

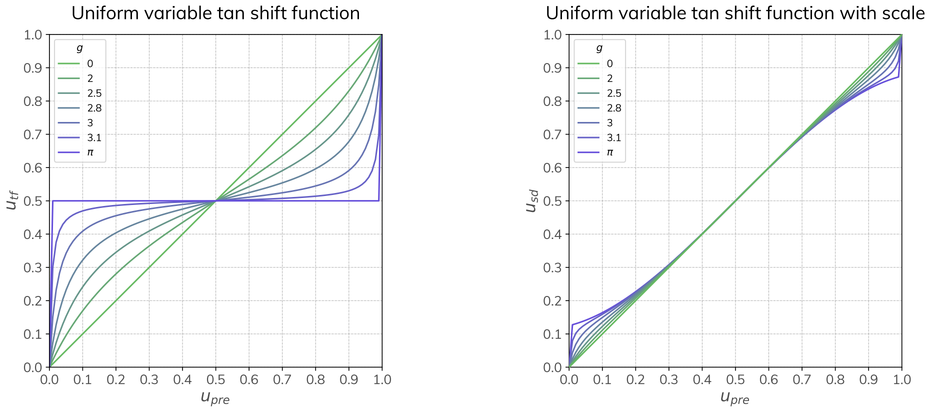
6.6.1. Shift samples outwards

This research poses a shift of the uniform variables in the fitted CDF, after a CDF is fit to the variable. The shift should make values in the extremes of the range [0,1], a little bit less extreme: for example, an extrapolation range of [0.99,1) could better be shifted to [0.97,1). This way, three times more samples can be drawn out of the extrapolated areas. This makes the model more likely to extrapolate, but it also makes the CDF less directly based on the data.

This shift is only based on the uniform value and prior uniform values of 0 should also be 0 in the shifted value, and similarly for 1, as otherwise not all uniform values (u) can be translated to a value (v). The tangens is a function that enlarges these ranges, when scaled correctly. When applied to the uniform space of a copula marginal distribution, the way to set up a tangens-based shift for the uniform value u_{tf} is the following:

$$u_{tf} = \frac{\tan(\mathbf{g}(u_{pre} - 0.5))}{2 \cdot \tan(\mathbf{g}/2)} + \frac{1}{2}, \tag{6.13}$$

where u_{pre} is the initially calculated uniform value u and $\mathbf{g}$ is a scale factor.



(a) Equation (6.13) for different values of g.

(b) Equation (6.14) for different values of g.

Figure 6.4: Shift functions

In Figure 6.4a, an example can be found of the shift function, for different values of the scaling factor g. This alters the middle section as well, and as this does not involve the part which can be extrapolated, this is unfavourable.

6.6.2. Limit shifted range

To solve this, a balance between no translation and Figure 6.4a is implemented. To focus largely on the edges, the factor is taken to the quadratic:

$$u_{sd} = u_{tf}(u_{pre} - 0.5)^2 + u_{pre}(1 - (u_{pre} - 0.5)^2). \tag{6.14}$$

In Figure 6.4b, Equation (6.14) is plotted for different values of g. The shift happens almost only at the edges, increases closer to the end and comes back to the points (0,0) and (1,1) to make a range that accepts all inputs. The total difference is also way less pronounced.

6.6.3. DIFFERENTIATE IN EXTRAPOLATION SIDES

For some variables, mostly an extrapolation from the upper limit is needed, or the other way around. For example, when the method keep underestimating high discharge peaks. In this thesis, this is the case (see Figure 6.7). Therefore, a shift can be made, to focus the change more on one half of the CDF. To do this, the 0.5 to get the shift of the CDF in the middle between 0 and 1, can be parameterised:

$$u_{sd,\mathbf{h}} = u_{tf}(u_{pre} - \mathbf{h})^2 + u_{pre}(1 - (u_{pre} - \mathbf{h})^2), \tag{6.15}$$

with **h** as the parameter.

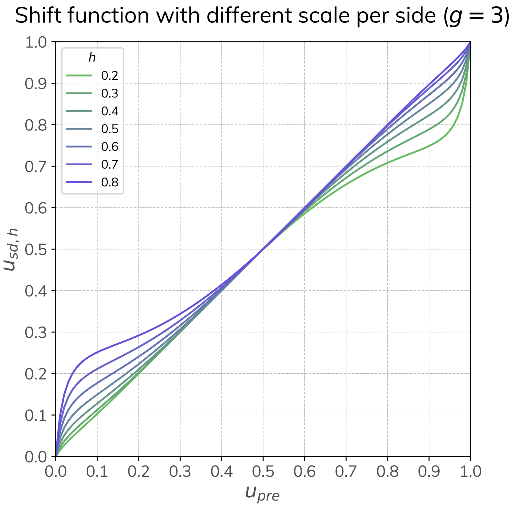


Figure 6.5: Equation (6.15) for different values of **h**.

An example of different values of **h** can be found in Figure 6.5. It shows that a bigger shift in one side holds a smaller shift on the other side, when done with **h**. Figure 6.6 shows a possible effect on the soil moisture CDF, as an example. **h** is 0.4, so there was an emphasis on the higher level. The figure shows clearly that the changes of high soil moisture levels are higher. For example, the chance of a soil moisture higher than 0.40 m^3/m^3, was initially about 0.02, but after applying Equation (6.15), it was approximately 0.10. However, the real extremes (> 0.45 m^3/m^3) still very rarely happen, as the function changes its slope to come back to (1,1) at the very end. This way, almost no highly implausible events (> 0.6 m^3/m^3 for example) are drawn.

6.6.4. IMPLEMENTATION

For the model used up to here, Equation (6.15) was implemented on all of the variables, to get better extrapolation results. For all of the variables, the was **g** = 2.65, as higher values distort

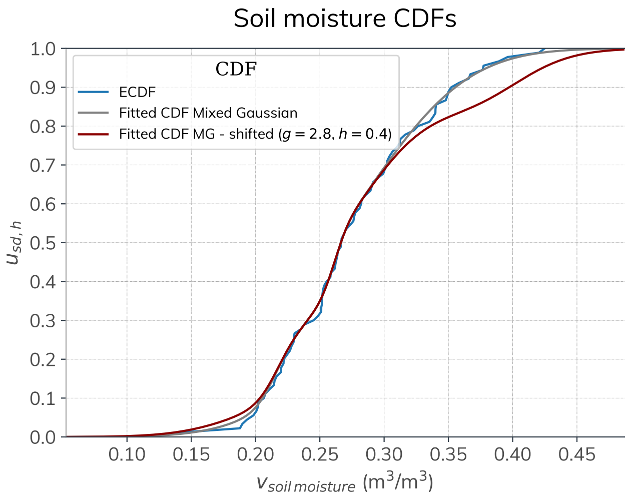


Figure 6.6: ECDF, fitted mixed Gaussian CDF and an example of a shifted CDF by Equation (6.15), of the soil moisture variable.

the distributions too much, and equal on both sides with **h** = 0.5, except for the MMDAD, as relatively high MMDADs were often predicted as too low, but these have the highest chance of creating a flood. Therefore, a **h** of 0.4 was used for this variable.

In Figure 6.7, the effect is shown on the error per discharge amount. The figure shows that the method shows less underestimation for the observations in the range higher than 9 m^3/s. For the middle section (5 to 9 m^3/s), the error increases, as the method overestimates more, on average. For the lowest discharges, the errors do not change significantly. There was a slight increase in median KGE for the model by using this shift: it increased from 0.72 to 0.73[3]. The decision on using the shift proposed, is therefore a choice between two alternatives that have an advantage and a disadvantage. As this research is interested in discharge extremes, it is chosen to implement this the shift as mentioned in this section.

[3] Tested with a 5-fold test, repeated 20 times for both implementing the shift and not using it.

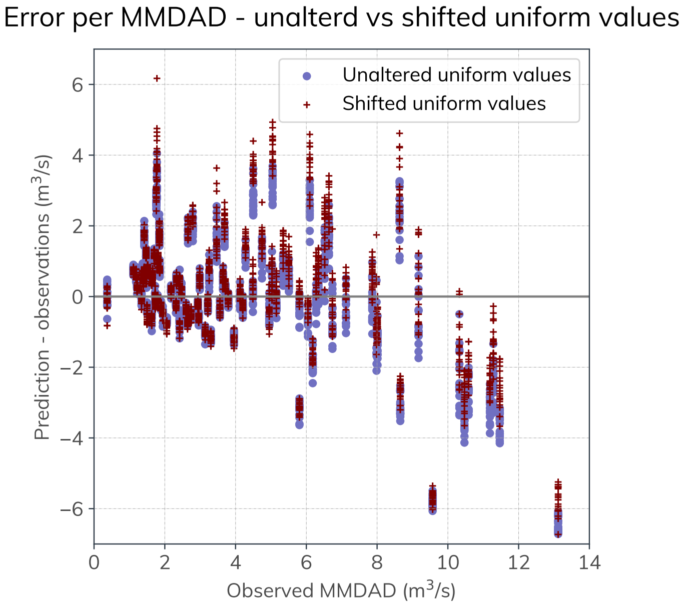


Figure 6.7: Error for different observations, prior and after applying Equation (6.15) with $g = 2.65$ and $h = 0.5$, except for the MMDAD variable, where $h = 0.4$. The shift does not make the model perform considerably better in general, but it differentiates in where the errors are: for higher discharges, the error decreases, whereas for median discharges (5 to 9 m^3/s), the error increases.

7

Model parameters and error sensitivity

Next to the optimal copulas to use and the optimal cumulative distribution function (CDF), there is a variety of other parameters that can be tested and changed for the Bayesian network (BN) model, to make the model perform better. This chapter tests these parameters one after the other, in which the model of the next parameter test uses the preferred results of all the previous tests. Testing all parameters at once through a Monte Carlo method would create calculating times in the order of days and is therefore out of the scope of this thesis. The tests are conducted by calculating the Kling-Gupta efficiency (KGE) for the BN predicting the monthly maximum daily average discharge (MMDAD), through a repeated k-fold test.

7.1. Most likely value from sampled values

The target variable is sampled n times (see Section 7.2 for the optimal value of n) to see what potential outcomes might be. This gives a probability distribution. However, for the ease of testing a parameter or model, practical use of the model, and communication of model results, a single value of the most likely outcome is useful. There are two easy ways to do this:

(i): Using the mean of the values. This is the most common method to calculate an expected value in the model. This has as an implication that, for example, when really high peaks are possible, the mean also shifts upwards and vice versa. Therefore, this might say more about the whole distribution than a single high possibility outcome. This is also the method that has been used up to this point in the research.

(ii): Using the median of the values. This means that there are exactly as many values that are higher, as there are lower than this number. This makes practical sense: this value has an equal chance of being too low as being too high. Moreover, when plotted with confidence intervals (see for example the figure in Section 9.1), this is a consistent, as both the confidence interval and the median are in a sense quantiles of the dataset. For example, the boxplots in Chapters 6 and 7 also use the medians as single value to show the most likely KGE per fold.

As both show advantages, it is also interesting to see which method delivers the best fit to

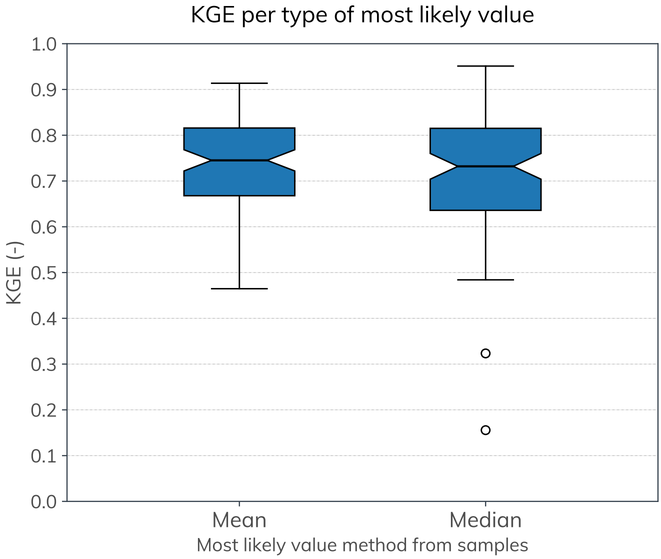


Figure 7.1: A comparison between the KGEs of using the mean versus the median of the samples. In general, using the mean gives results that are slightly closer to the actual values.

the measured values. Therefore, the model that is acquired from the optimal CDF in Chapter 6, is tested for both of these methods. It is tested with k-fold with $k = 5$, and repeated 20 times with other random folds to get a good average. This sums up to 100 folds and KGEs per method. These results can be found in Figure 7.1. The plot shows that using a mean of the samples delivers a slightly better fit to the data. Moreover, the results from using the median are a little less stable, with two folds being outliers and the range of 50% of the data (the dented box) being a little wider. For this research, this is the decisive argument to use the mean. However, it is possible to use the median when the arguments posed in (ii) are more important, as this does not deliver dramatically worse results.

7.2. NUMBER OF SAMPLES TO USE

The copula is sampled according to Section 2.6. It is interesting to see how many samples are needed to acquire sufficient predictive values. There are three conditions discussed in this research for the number of samples to use.

7.2.1. GOOD FIT OF THE TARGET VARIABLE

To test how many samples are sufficient to get a good fit of the target variable, the model is tested 5-fold with 10 repetitions of random folds, so 50 folds and KGEs in total per number of samples. The results can be found in Figure 7.2. The figure suggests that using 20 samples would already be sufficient to get results that are similar to results from using a higher number of samples.

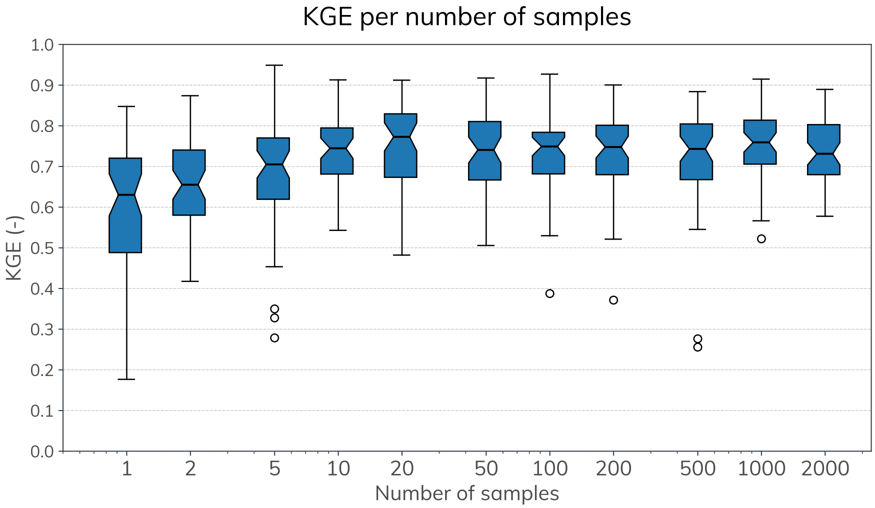


Figure 7.2: Boxplot of the KGE per number of samples used to predict the target variable from the Gaussian copula. The performance seems to level around 20 samples onwards.

7.2.2. CONFIDENCE INTERVALS

When confidence intervals are required to get a sense of the certainty of the calculated value, the number of samples should be sufficient to calculate stable quantiles. For example, when of 100 samples, a confidence of 90% is used, it leaves 5 values of the data on each side. As variable usually have a low probability towards the edges of the distributions, especially since they have been sampled from a Gaussian copula, these 5 values have a higher volatility than the values in the centre. That is why this does not deliver a stable number of samples. The rule-of-thumb that is followed in this research, is to use at least 50 samples in each of the margins outside of the confidence interval. As the confidence interval used in Section 9.1 is 80%, at least $\frac{50}{0.1} = 500$ samples are necessary.

7.2.3. COMPUTATIONAL TIME

In the testing of the model, most of the time running the model goes into fitting the CDFs. The computation time only increases noticeably when testing with more than 10,000 values in the Python code written for this research. Below this number, the calculating time was in the order of 7 to 9 seconds per run but for more than 10,000 samples it increased to more than 10 seconds for most of the folds. From Sections 7.2.1 and 7.2.2 it follows that this number of sample values is not necessary to acquire good results. Therefore, as the computational time is no limiting factor for the rest of this research, no lower number of samples needs to be used, and the number of 500 samples is used from this section onwards. However, when only getting results from a model that has a wide confidence interval, many predicting datasteps, and potentially intermediate results, computational speed might be a limiting factor.

7.3. SUBSET PERIOD DATA

The data for the model has been aggregated from the days before the MMDAD event, see Section 3.3.9, step 1. As the model is almost solely data driven, it is unclear at the start which period to take into account. Therefore, a Monte Carlo test has been done with a 10 time, randomly repeated 5-fold test of the model for 9 different periods before the maximum event. For example, when this subset period was 4 for a MMDAD on June 8, the data was taken from June 4, 5, 6 and 7 and the first 12 hours of June 8.

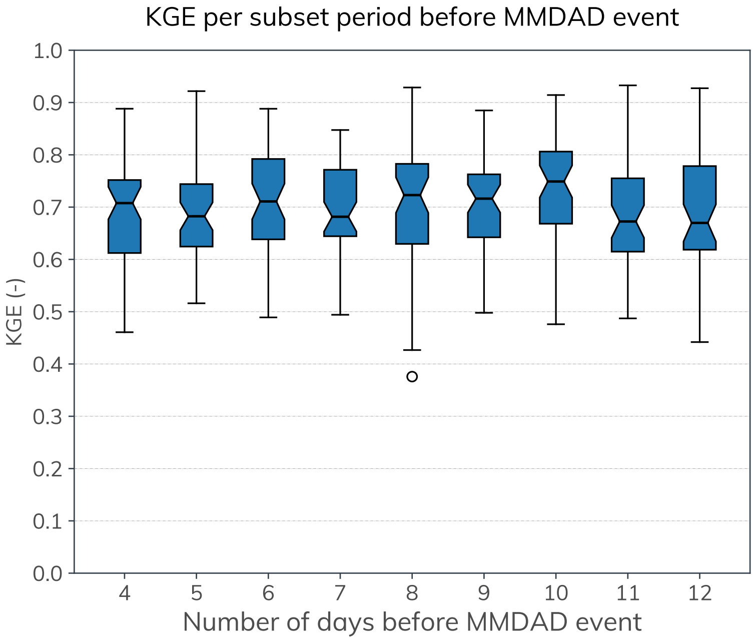


Figure 7.3: The amount of data days before the discharge event of which the data is aggregated to construct the model.

It is clear from Figure 7.3 that it does not matter greatly what subset period to use. 8, 9 and 10 days back seem to perform the best, but they do not significantly outperform using anywhere between 6 to 10 days as a subset period. Therefore, in this thesis, a subset time of 8 days was taken.

7.4. SENSITIVITY INPUT ERROR

It is common to recognise two types of errors, systematic and random errors. For this research, it is also interesting to look into whether the error is in the conditioning variables, or in the conditioned variable(s).

7.4.1. Systematic errors or bias

A continuous systematic error, or even an error that is a direct monotonically increasing function of the actual values, does not influence the BN results. This is because of the translation to the uniform margins for the copulas. What is interesting, however, is to test what the influence of a new systematic error is.

Multivariate normal function

When a variable changes by a parameter, the vector A from Equation (2.18) is shifted with ΔA, which is the shift in standard normal values per conditioning variable. This means that the new mean of the conditional normal distribution subject to an error, $\hat{\mu}_e$ is defined as:

$$\hat{\mu}_e = R_{12} R_{22}^{-1} (A + \Delta A)\,. \tag{7.1}$$

The shift in mean is then given by:

$$\Delta\hat{\mu}_e = R_{12} R_{22}^{-1} \Delta A. \tag{7.2}$$

However, the implication on the predicted variable depends on both $\hat{\mu}$ and $\Delta\hat{\mu}_e$, as $F(v)$ from Equation (2.3), as well as Equation (2.20) are non-linear functions. This also holds for ΔA. Therefore, to see the shift in the prediction for this research, a simulation with artificial systematic errors is conducted.

Simulation artificial errors

There are 7 conditioning variables that could have any or no systematic error. Each of the variables has been tested with an artificial systematic error, one after the other, for its influence on the prediction of the MMDAD. The error is a factor times the difference between the highest and the lowest value. This factor is called the relative error (ϵ_r), and is added to the values of a variable. There is no k-fold test done, as the test set is already altered by the error. The results can be found in Figure 7.4.

The results show that an error in variables that are not directly correlated to the MMDAD, has no significant impact on the final results, as they have very little influence whatsoever when the child variable's value is known. However, for the other variables, the impact can be greater. From an absolute, relative error of 0.5 times the range of the values for a single variable, the results start to become unsatisfiably worse. However, all of the direct variables also have a region in which the prediction actually gets better. For the negatively correlated variables of surface water level and solar radiation, this happens when small values are subtracted from the measurements, and for the positively correlated variables this is the other way around. A possible explanation for this, is that the variables are averaged too much in the BN, such that insufficient extremes are predicted.

7.4.2. Random errors

Theory

In this research, a continuous standard error is assumed to influence the measurements (σ_m). The influence of these errors on the the network, can first be studied by looking at the error on the standard error of the subset values. In mathematical terms this is called the standard error of the means, and is defined as follows:

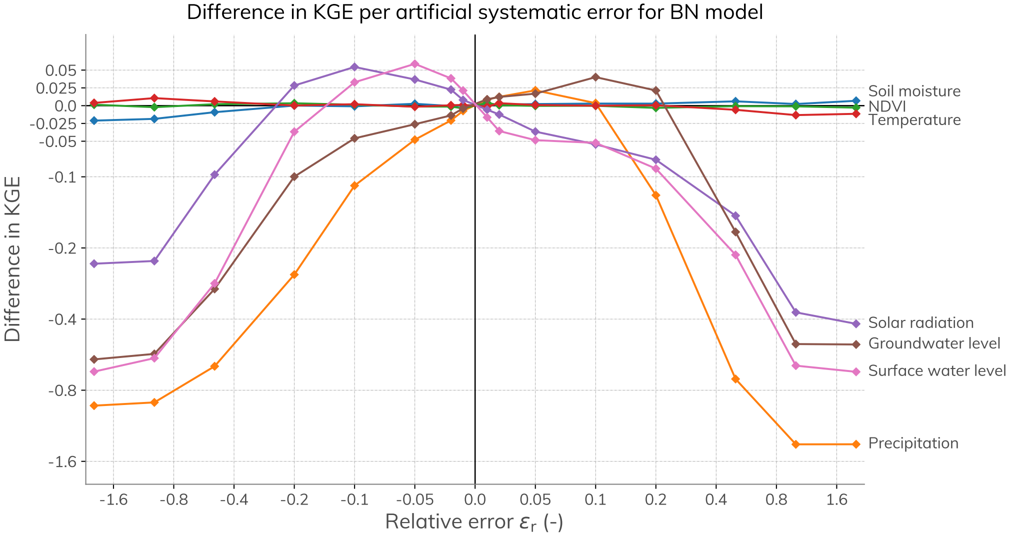


Figure 7.4: The loss/gain in KGE for different artificial errors. Note that both axes are logarithmic divided by 10, both for the positive value and the negative. This is why between -0.1 and 0.1 for the x-axis and -0.2 and 0.2 a shift in scale is made by applying a linear axis (symlog axis).

$$\sigma_v = \frac{\sigma_m}{\sqrt{n}}, \tag{7.3}$$

where n is the number of measurements in the subset period. For the subset period of 8 days, the standard error of the value is calculated with an artificial standard error for the measurements 0.2B, where B is the range between the highest and lowest measured value of a variable. The results can be found in Table 7.1.

Variable	**Avg. number of measurements in subset period (n)**	**Standard error of means with $\sigma_m = 0.2B$ ($\times B$)**
MMDAD	96	0.0204
Precipitation	204	0.0140
Temperature	204	0.0140
Solar radiation	204	0.0140
Soil moisture	9	0.0667
NDVI	1.495[1]	0.164[1]
Groundwater levels	119	0.0183
Surface water levels	1080	0.00609

Table 7.1: Standard error factorisation of means

[1] NDVI data is already aggregated by NASA, so a σ_m of 0.2 would be way more significant for these measurement. It is to be expected that the standard error of this measurement is lower than that of real individual measurement.

It is clear that the error of the subset periods increases significantly, as most variables have many values. For the soil moisture, however, the error remains significant. The data of the NDVI is already aggregated by NASA, so it does not compare proportionately.

SIMULATION ARTIFICIAL ERRORS

To test the influence of an additional random error, the model has been tested with an artificial random error. This is composed as follows: per variable, for each of the timesteps, there is a random normal error added to the value. The errors are assumed to follow a Gaussian distribution, so this random error is drawn out of $\mathcal{N}(0,\sigma_v)$, where σ_v is defined according to Equation (7.3). In this function, for n the average number of measurements is used (see Table 7.1), and σ_m is expanded as:

$$\sigma_m = \epsilon_{r,r} B, \tag{7.4}$$

where B is the largest value of the variable minus the lowest value (max(V)-min(V)), and $\epsilon_{r,r}$ is called the relative error, a factor that is changed in this method. This procedure is done 20 times per variable per $\epsilon_{r,r}$ to get a good average. The results can be found in Figure 7.5.

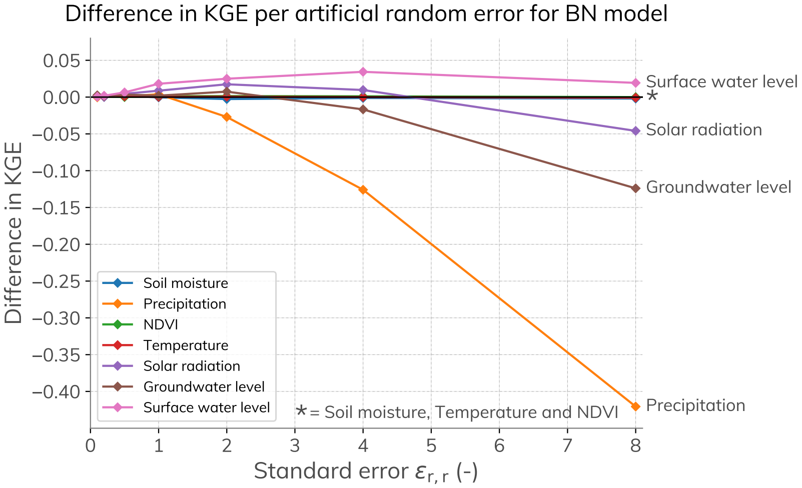


Figure 7.5: Artificial random error and the influence on the influence in performance in predicting the MMDAD. The factor $\epsilon_{r,r}$ is defined as σ_m/y, or in other words, the standard error is the factor $\epsilon_{r,r}$ times the range between the highest and the lowest value of a variable (B).

According to the test, for almost none of the variables, a random error has any significant impact on the results. A relative error of 4, which is already very high, makes no impact on the variables that are not directly correlated and very little in the results of the surface water level, solar radiation and groundwater levels. As for the surface water level and solar radiation, an increase was even visible, which is remarkable and suggests that these variables maybe not be optimal predictors. Only the precipitation already has a large decrease in KGE from an $\epsilon_{r,r}$ of 4 onward. In conclusion, the model is very robust against random errors in the data.

7.4.3. SYSTEMATIC ERRORS IN THE DISCHARGE MEASUREMENTS

If the discharge measurements were to have systematic errors, the model is fitted wrongly. However, the predicted discharges would be the same as if the model would have been trained by the correct measurements, shifted by the the systematic error in the discharge measurements. For the KGE, this means that the factors r and α do not change in Equation (4.2). The μ_{obs} could be redefined as follows:

$$\mu_{\text{obs}} = \mu_{\text{act}} + \epsilon_{\text{s},MMDAD}, \tag{7.5}$$

where μ_{act} is the mean of the actual measurements and $\epsilon_{\text{s},MMDAD}$ is the systematic error in the observations.

If the actual discharges would be known, it would be possible to calculate the KGE with μ_{act} instead of μ_{obs}. This gives a sense of the sensitivites in the KGE in the case of potential systematic errors.

In the BN proposed in this research, the mean of the simulations are almost equal to the mean of the observations (see Section 9.1.2). This means that the factor β (Equation (4.3b)) in the KGE formula (Equation (4.2)) is almost 1. In this section, this factor is presumed to be exactly 1, yielding the following difference in KGE (ΔKGE):

$$\begin{aligned}\Delta KGE &= 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + \left(\frac{\mu_{sim}}{\mu_{\text{obs}} - \epsilon_{\text{s},MMDAD}} - 1\right)^2} - KGE_{\text{pre}} \\ &= 0.28 - \sqrt{0.28^2 + \left(\frac{\epsilon_{\text{s},MMDAD}}{\epsilon_{\text{s},MMDAD} - 4.24}\right)^2},\end{aligned} \tag{7.6}$$

where r, α and μ_{sim} are the usual KGE variables, μ_{obs} in β has been substituted for μ_{act} , with the latter defined in Equation (7.5), and KGE_{pre} is the original KGE. The second line of Equation (7.6) is obtained by assuming that $\beta_{pre} = 1$, and noting that KGE_{pre} was 0.72 on average. The full derivation can be found in Equation (C.5) in the appendix. Equation (7.6) has been plotted in Figure 7.6.

From the graph it follows that, in general, positive systematic discharge errors are worse for the KGE than negative systematic discharge errors of the same magnitude. The reason for this is that in Equation (7.6), the magnitude of the delimiter becomes smaller when $\epsilon_{\text{s},MMDAD}$ is larger than 0. This does not mean that it is in all cases better for the model to have underestimations in comparison to overestimations, as the KGE is just a measurement of performance. The users' view on the performance, is more related to the goal of the model. If, for example, the model is made for flood safety, it might actually be better to measure a higher discharge than which is actually happening.

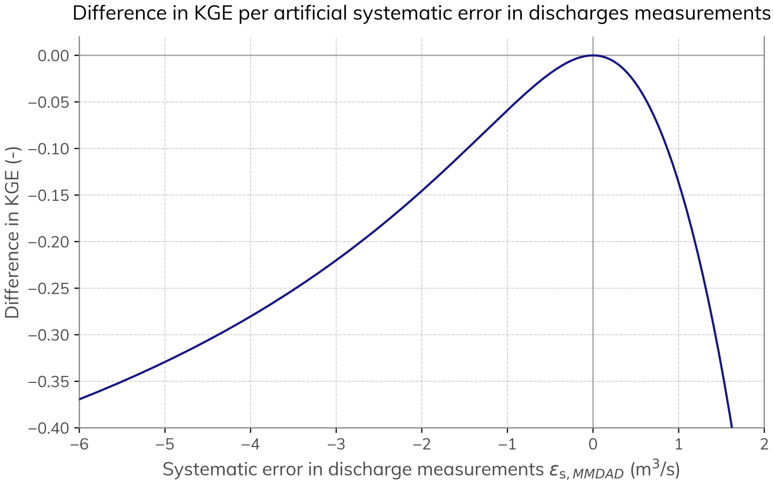


Figure 7.6: The result on the KGE if the discharge measurements where to have a systematic error and the model was benchmarked against the actual discharge.

8

BAYESIAN NETWORK LAYOUT

Up to this point in the research, sensitivity analyses and optimisation steps, where only of the parameters and settings related to the copulas. For the user of the model, the visible part of a Bayesian network (BN) is the most intuitive and comes closest to reality. Furthermore, the layout of the BN has considerable impact on what can be conditioned and how well some variables can be predicted. It also can influence whether influences of a certain variable can be distinguished correctly. For example, if it is known that on a certain date, the temperature will be higher, you might want to see what the influence on the network of this variable is. If this mostly influences the precipitation, which influences the rest of the variables, the influence of the effect of the precipitation is mostly seen. This chapter goes into the creation of a BN layout, that suits the catchment best and achieves the the modelling goal of this research.

To come up with the preferred layout for this research, first the criteria what make a layout superior to others, have to be drawn up. Afterwards, a strategy is composed to acquire these criteria, based on the criteria and the characteristics of non-parametric Bayesian networks (NPBNs). When applying the strategy, some implementations are ambiguous: implementing a section in a certain way or another way, explains the catchment workings differently. As the model is no perfect representation of reality, neither argumentation is wrong. In this research, the choice what implementations to take, is solely based on performance of certain predictions in the model, as no other method can be used to make an unambiguous decision.

8.1. CRITERIA

In this research, the following criteria are used to assess the BN layout[1].

1. **Physical relations:** the model layout should adhere to logical cause-effect relations. This makes the model physically relevant and better interpretable.

2. **No useless variables:** the model should not contain variables that do not add any predictive power to any of the other variables. This keeps the model relevant and orderly.

[1] See Appendix A in the appendix for an explanation of the terminology used in this section.

3. **Accurate predictions and no overfitting:** the model should perform well, even when tested on data that is not in the training data.
4. **Understandable model:** the model should be understandable for anyone that has a decent amount of proficiency in hydrology and has rudimentary knowledge about BNs.
5. **Variables predictable from other variable when not known:** in case the value of a variable is not known at a certain timestep, the other variables should update the distribution of that variable in line with other datapoints. The updated distribution should also help make the update of other variables - such as the target variable - better comply with observations.

An additional condition is that the configuration cannot have circular relations (that the same point in the graph can be reached by following arcs in their direction) as BNs are directed, acyclic graphs (DAGs).

8.2. Strategy

The following strategy to create a BN layout that complies with these criteria has been established:

1. Select potential variables to use in the model.
2. Calculate normal rank correlations for all combinations and remove variables that do not hold normal rank correlations higher than 0.3 with any of the other variables and are therefore not of use to make predictions.
3. Use physical relations to create a model.
4. Determine order of dependencies to child variables.
5. Remove connections with low partial normal rank correlations from the BN (<0.1). Usually, these connections do not make a significant influence on the update of distributions of other variables. An exception can be made to this rule if the the connection does make a noticeable difference in the update of variable which is regarded as important.

The normal rank correlations of the variables used in this research can be found in Figure D.8 in the appendix. None of the variables first introduced have to be removed. Despite this, the low correlation of the precipitation measurements is noteworthy.

8.3. Connections in the model based on influences on the variables

To solve Items 3 and 4 of Section 8.2, a look is taken at which variables could potentially influence other variables in the actual catchment. The former variables will be te parents of the latter variables. The following list goes through all of the variables and determines which variables could have influenced that variable. Then, it discusses the order of dependencies per variable. For some of the arcs, multiple interpretations of the system are possible. Therefore, it is merely a choice of which setup to implement. In this research, the corresponding decisions are solely based on accuracy of the model, see Section 8.4.

- **Solar radiation:** The solar radiation is determined by the solar incidence angle, cloudiness, and other minor atmospheric parameters. These other atmospheric parameter are not known, so they cannot be implemented. The cloud cover could be correlated to the precipitation. Therefore, this variable has a small potential to be a parent. However, the normal rank correlation between precipitation and solar radiation is only -0.14 (see Figure D.8), which is very low. Therefore, the solar radiation does not have any parents in the model.

- **Precipitation:** Due to the climate in The Netherlands, the precipitation pattern is very similar throughout the year (see Figure 3.6). Therefore, a season-based variable, such as solar radiation or temperature, is not a good candidate as a parent variable. Moreover, rain clouds originate almost fully from outside the catchment area (in contrast with large rainforest catchments), which gives an indication that the precipitation is likely not influenced by any other variable. Therefore, precipitation is also regarded as a top-level variable.

- **Temperature:** The temperature is influenced by solar radiation, weather factors, and local solar radiation reflection and cooling effects. As the area of this catchment is relatively small for local weather influences, and the terrain is relatively flat, local effects are presumed to be insignificant. Solar radiation influences the temperature directly and is therefore implemented this way. The feedback mechanism through evaporation and cloud formation is regarded as insignificant to this direction.The same could be said for using NDVI, because of the cooling effect of plant transpiration. This is choice (1).

  When the NDVI is a parent of the temperature, its correlation is conditional to that of the solar radiation and the NDVI, because the solar radiation influences the plant activity as well.

- **NDVI:** This vegetation index is a proxy for leaves cover and plant activity in the catchment. As plants need solar radiation to grow, this is the first parent for the NDVI. Regarding the connection with temperature: plants also fare better with a high temperature, thus the arrow between the NDVI and temperature could also be the other way around in choice (1). Finally, soil moisture availability also influences the amount plants can grow. This function is physically not monotonic, as there is a strict margin (field capacity and wilting point) within which the plants can grow. Therefore, the NDVI is influenced by the soil moisture rate.

  Unconditionally influencing the NDVI is the solar radiation. If implemented in this fashion, the temperature comes second, and is only conditional to the solar radiation. The soil moisture, if used as such, is dependent on the other two variables to remove the seasonal effect from it and let the soil moisture explain the relative water availability at that moment.

- **Soil moisture:** As the soil moisture measurements are of the top soil, this consists mostly of unsaturated soil moisture. This quantity is mostly influenced by the precipitation, suction of plants, capillary rise from groundwater, or just directly from groundwater rise (because of seepage or saturation effects) if this reaches the top level of the soil. Moreover, evaporation forces (temperature, solar radiation) can also influence soil moisture.

The first parent variable is the precipitation, as this is presumed to be very directly causing soil moisture differences. Secondly, arguments can be made for connecting groundwater to the soil moisture in either direction: this depends on whether capillary rise or percolation is the dominant process. This is choice (2). Lastly, it is debatable how much evaporation happens in the catchment in relation to transpiration. This is choice (3).

Whether to select the groundwater or the precipitation as first variable for the partial correlation, is also a choice (4). After this come the solar radiation and the temperature, in that order, if they are used.

- **Groundwater level:** As mentioned above, soil moisture has one of the largest effects the groundwater level, to be selected in choice (2). Groundwater is depleted again via the surface water, and is afterwards discharged at the downstream end of the catchment. However, arguments can also be made that the surface water actually influences the groundwater level. This is because the surface water level is located in a managed area (see Figure 3.2), in which the surface water level is artificially raised to create a higher groundwater level. However, as can be seen from the same map, the measurement stations are also relatively far away from each other, which can also be an argument for there not being an arc in the BN. This is choice (5).

  If surface water level is a parent variable to the groundwater, this will have the unconditional correlation to the groundwater, as they are very closely related in managed areas. The influence of the soil moisture is then used indirectly: given the surface water level.

- **Surface water level:** As mentioned in Section 3.2.6, the area of surface water is only 1.36% of the catchment. This means that the influence on direct precipitation on the surface water is low in comparison to the groundwater, which is the main contributor. Therefore, choice (4) determines what to do with the surface water levels. No other variables are connected. This means that the surface water level will never have multiple parent variables, so no choice is to be made in order of partial correlations.

- **Monthly maximum daily average discharge (MMDAD):** The discharge does not influence anything in the catchment, as it (almost always) flows out of the catchment. What influences the discharge is, on the other hand, up for interpretation. This is choice (6). Directly, it is only the downstream surface water level that influences the discharge. However, the measurement station is only in a small tributary, and using a single variable does not make for a good prediction. That is why, in this research, also the groundwater level, precipitation and the soil moisture (that are more directly possible to relate variables than others) are used to predict the discharge directly. Lastly, also a variable with a strong predictive power of the season is added: the solar radiation.

  The order of dependence between the surface water and groundwater levels is determined for choice (5). As the direct relation of some of the other variables is more or less imaginary, no strong arguments can be made for the partial correlation order of the other variables. In this research, the order is: surface water level, groundwater level, precipitation, soil moisture, and solar radiation.

8.4. SELECTED IMPLEMENTATIONS

It is not possible to make a completely satisfying decision on the BN layout, because choices have to be made between implementations that both have positive and negative aspects. In this research, we will therefore decide upon these choices solely based on the performance of predicting a variable in the model. Users of similar models can make other decisions, based on their preferences and other leading processes in catchments.

All the choices with a small number of variables (choices (1) to (5)) are recapitulated in Table 8.1. Selecting a certain implementation in choice (6) is very arbitrary, and (given that the surface water level is connected to the MMDAD) still has $2^6 = 64$ different options. However, all of the different options in models (1) to (6) still have $3 \cdot 3 \cdot 2 \cdot 3 \cdot 2 \cdot 3 = 324$ different implementations, which is too time-consuming and labour-intensive for this research.

Choice		**Implementations**		
Num.	Connection	A	B	C
1	NDVI - Temperature	No connection	NDVI → Temperature	NDVI ← Temperature
2	Groundwater - Soil moisture	Groundwater → Soil moisture	**Groundwater ← Soil moisture**	
3	Solar radiation & Temperature - Soil moisture	No connection	**Solar radiation → Soil moisture**	Solar radiation & Temperature → Soil moisture
4	Groundwater & Precipitation - Soil moisture	Groundwater dependent: $\rho_{Prec,SM}$, $\rho_{GW,SM}\vert\rho_{Prec,SM}$	**Precipitation dependent:** $\rho_{GW,SM}$, $\rho_{Prec,SM}\vert\rho_{GW,SM}$	
5	Surface water level - Groundwater level	No connection	**Surface water level → Groundwater level**	Surface water level ← Groundwater level

Table 8.1: Overview choices model. In **bold** are the implementations that are used in the intermediate models to solve the choices before this choice.

Therefore, each of the choices is regarded one after the other, from (1) to (6). For each of these choices, a test is constructed to test the performance of the model in a specific section of the model. The model is only conditioned on the variables that are of interest for the choice at hand. The reason that this method is chosen, instead of conditioning the whole model and testing the accuracy of the MMDAD prediction (in Kling-Gupta efficiency (KGE)), is because no high differentiation in performance is expected, especially for variables that are not direcly connected to the MMDAD. Moreover, the model should also perform well in predicting other variables.

Each of the different implementations was tested 5-fold, with a 20 times random repetition, hence providing a total of 100 results per implementation. For each of the implementations, the median of the KGEs was taken by comparing the target variable predictions and observations. This is done because in very few cases, the model gave very inaccurate results and gave KGEs in the order of -10 to -60, which has a big influence on the mean of the KGEs. The optimal implementation was chosen as the one with the highest KGE and this setup was used in the testing of the other choices from here. The results of these tests can be found in Table 8.2.

Choice	Model layout		Median KGE per impl.			Sel. impl.
	Target	Conditioned	A	B	C	
1	Soil moisture	NDVI, Temperature	0.26	0.23	0.24	A
2	MMDAD	Soil moisture, Groundwater level	0.34	0.38	-	B
3	NDVI	Solar radiation, Temperature, Soil moisture	0.37	0.38	0.41	C[2]
4	-	-	-	-	-	-[3]
5	MMDAD	Groundwater level, Surface water level	0.41	0.42	0.39	B

Table 8.2: Outcome testing all of the choices an selected implementations.

Table 8.2 shows that the KGEs of the different implementations do not differ a lot. Therefore, the selected implementations should not be regarded as the overall optimal setup, but just the optimal model - by a small margin - for the KGE of a part of the model. Implementation C for choice 4 delivers a connection with a partial normal rank correlation of -0.089, so slightly less than absolute 0.1. This is in conflict with Item 5 of Section 8.2. However, as this arrow does add something to the prediction of the NDVI, the connection is kept in. As the result from choice (2) is an arrow from soil moisture to groundwater, the choice of what the order of dependencies between the groundwater and precipitation is, choice (4), is redundant.

Regarding the final choice (6), the variables surface water level, groundwater level, soil moisture, solar radiation and precipitation have been selected to be connected to the MMDAD.

8.5. Final model

The final model is shown in Figure 8.1. See Appendix A for an explanation how a BN layout works.

[2] This keeps an seemingly redundant connection in ($\rho_{s,norm} < 0.1$), however, this does make a better prediction. See explanation in text Section 8.4.

[3] Because of the implementation of choice (2), this choice has become redundant.

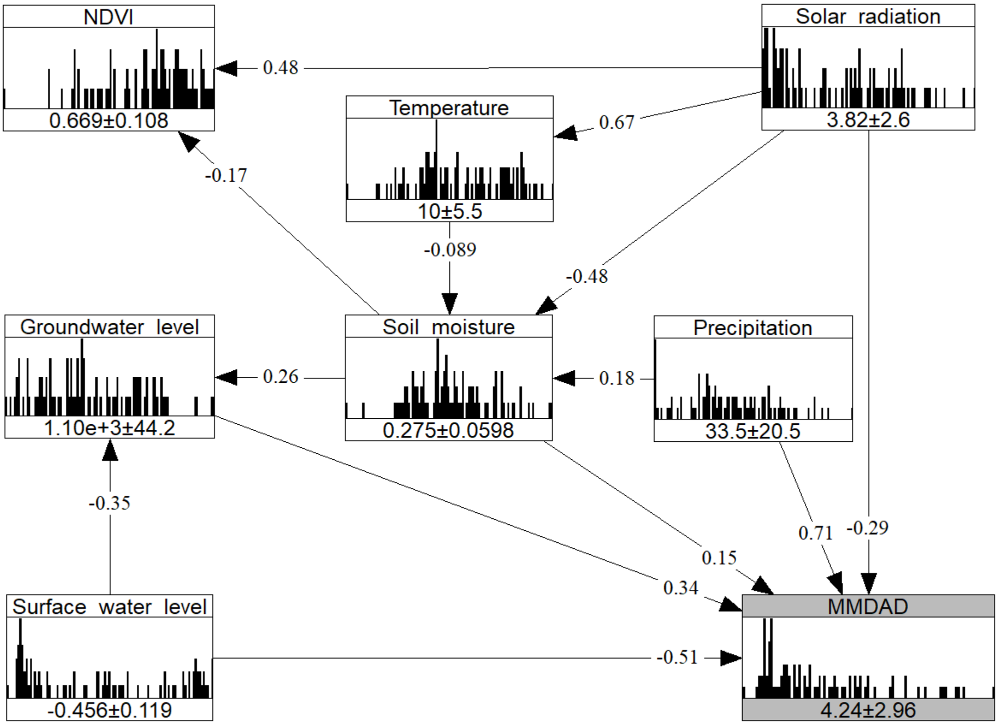


Figure 8.1: Final model layout. The numbers at the bottom of the boxes are the mean and the standard deviation of the values, in their respective units: NDVI (-), temperature (°C), solar radiation (J/cm^2), groundwater level (hPa), soil moisture (m^3/m^3), precipitation (mm, cumulative over the whole subset period), surface water level (m). The numbers on the arcs are the (conditional) normal rank correlations between variable pairs, defined by Equations (2.12) and (2.17). See also Appendix A for a more extensive description of a BN layout.

9

Results Bayesian network and benchmark models

In this section, the results of the final model, that has been optimised in the previous chapters, are discussed. Afterwards, a number of other models is introduced and tested with the data, to be able to benchmark the results of the Bayesian network.

9.1. General results Bayesian network

The median performance of the final model was a Kling-Gupta efficiency (KGE) 0.73 per fold.

In Figure 9.1 a model is made on all of the data in the dataset (and is therefore not k-fold tested). It can be seen from the model that most of the observations are within the 80% error bar. The absolute error is usually the highest for the highest peaks. This is common. However, as high peaks usually deliver the worst floods, it is also not favourable.

9.1.1. Complete k-fold test

In Section 4.3, it is discussed how only the fitting of the CDFs can be performed automatically in the Python code written for this thesis, via k-fold cross validation. As a final evaluation, a complete k-fold test has been executed, where also the matrix R (see Section 2.4.2) is created from the data in the training set via Uninet. The results of the 5-fold test that has been executed can be found in Figure 9.2. Evidently, the separate R-matrices do not lower the KGE significantly. The average difference between the maximum correlation and the minimum difference of the 5 folds, per connection, was 0.06. Additionally, the maximum of these numbers was 0.14. This was the connection between the monthly maximum daily average discharge (MMDAD) and the surface water level.

9.1.2. Factors KGE

As was mentioned in the previous section, the average KGE was 0.73. However, since the KGE is used, it is known how different KGE parameters have influenced this number. First of all, the correlation coefficient ρ between the predicted values and the observations was 0.85 on

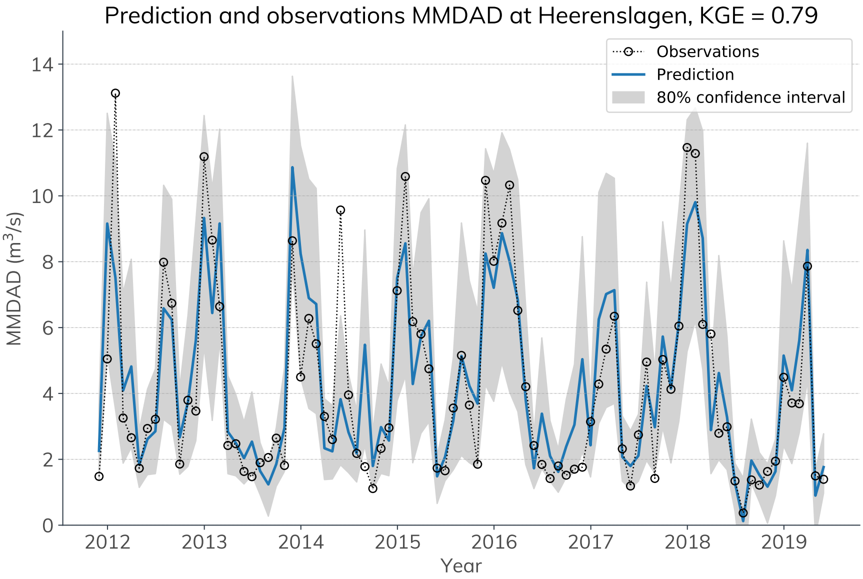


Figure 9.1: Prediction and measured observations at Heerenslagen. This specific model is not k-fold tested. cumulative distribution function (CDF): mixed Gaussian with 3 normal distributions, most likely value: mean (see Section 7.1), number of samples: 500, shifted CDF as mentioned in Section 6.6.4.

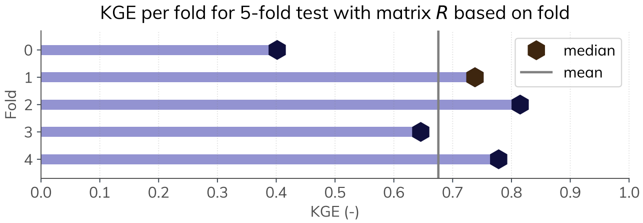


Figure 9.2: Result of a 5-fold test where also the matrix R has been calculated per fold. See Section 4.3.

average. This is one of the reasons for the small KGE. Secondly, the factor α, the difference between the two standard deviations, was on average 0.83. It can also be seen from Figure 9.1 that the volatility of the model is slightly lower than the observations. One of the reasons for the reduced volatility of the model, is the underestimations of the high peaks. Lastly, the factor β was on average 0.9997. This means that the mean of the predictions is about the same as the mean of the observations. This can be explained by the fact that a probability distribution is made from the observations. The mean of this distribution is also drawn out of the random samples, and the factor has about an equal chance of being higher than 1 as being lower. However, it is still remarkably close to 1 most of the times.

Additionally, the median Nash-Sutcliffe efficiency (NSE) of the final model was 0.66, when 5-fold tested for 20 times. Optimising the model for the KGE, might have not given the optimal model for the NSE. Both determine performance with a slightly different philosophy, which makes that the optimal model is optimised for the KGE philosophy.

9.1.3. Error per observed values

For 20-times 5-fold testing, the results have been plotted for the observed discharge versus the error (predictions minus observations) in Figure 9.3. It is clear that in general, for the low discharges, the error is often small, but many folds overestimate the discharge significantly. For the median discharges (5 to 8 m^3/s), the general error is higher, but this is still an overestimation. Finally, for the highest discharges (8+ m^3/s), the model tends underestimates the discharge slightly. Two observations in the high range are never really estimated well.

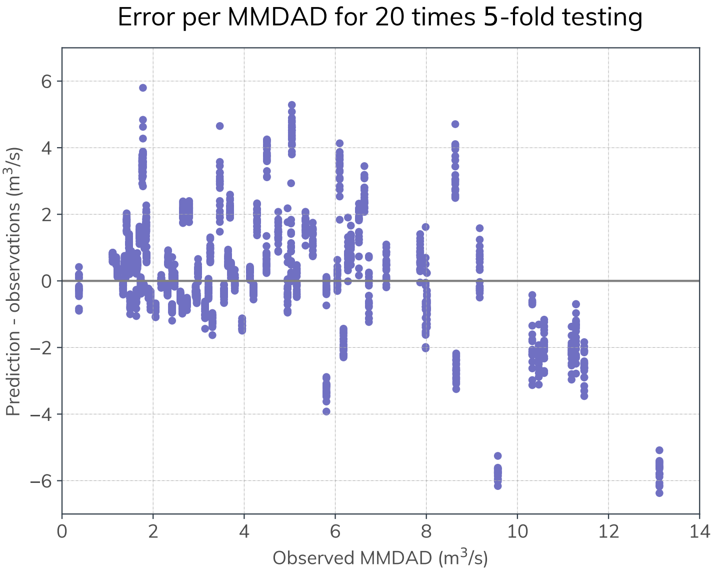


Figure 9.3: Model results for the final model, plotted for error versus the observed MMDAD.

9.1.4. Not fixing all variables

For almost all of the tests in this thesis, the target was the prediction of the MMDAD based on fixing all of the other variables. However, Bayesian networks are also able to condition the distribution of the target variable if not all other variables are fixed. If this is used, the non-fixed variables are also conditioned by the fixed variables, what works through in the network. In this section, the MMDAD is predicted by fixing a number of variables in the network. As there are 7 variables in the model except for the MMDAD, there are $2^7 = 128$ combinations. Testing all of these combinations, will give very little overview in the results. Therefore, in this thesis, 1, respectively, 2 variables are fixed. This gives a total of:

$$n_{\text{comb}} = \binom{7}{2} + \binom{7}{1} = 21 + 7 = 28 \tag{9.1}$$

combinations. For each of variables or variable combinations, a 5-fold test is repeated 10 times with different folds with the final model. In Figure 9.4, the median KGE can be found of the

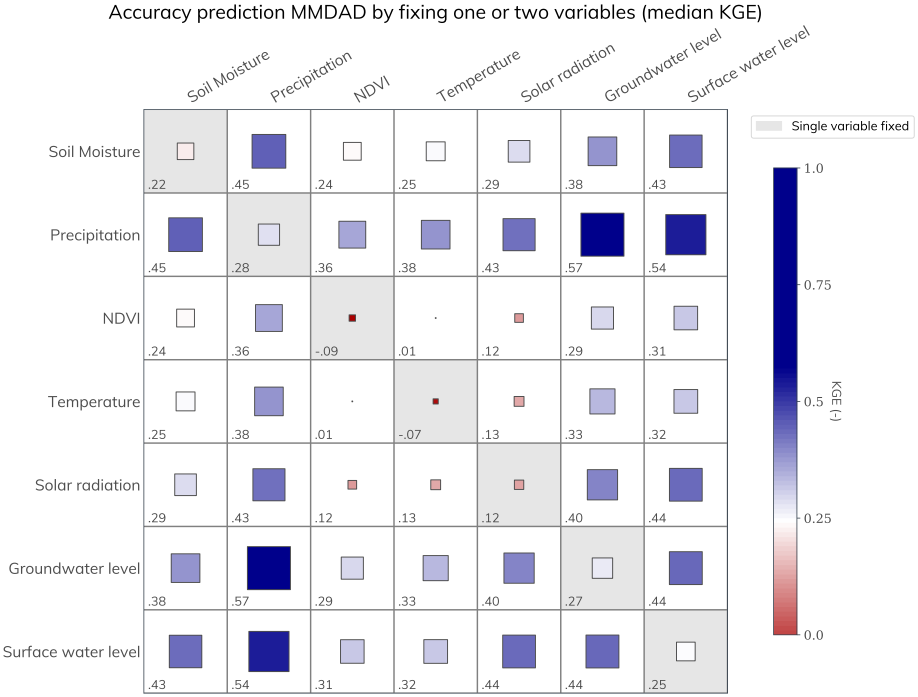


Figure 9.4: Median KGE for 5-fold test, repeated 10 times, predicting the MMDAD by fixing one or two variables. On left top to right bottom axis are the results where only one variable was fixed.

test. In general, the precipitation, groundwater level and surface water level are the best predictors in this model. However, combining precipitation with one of these water levels gives a significant better result than combining the groundwater level with the surface water level. After these variables, the soil moisture is a decent predictor. The NDVI and temperature on itself are useless predictors. However, combined with for example, precipitation, the combined prediction becomes significantly better.

9.1.5. PREDICTING ALL VARIABLES

As mentioned in Section 9.1.4, the Bayesian network update probability distributions of other variables (than solely the target of this research) as well. How well these other variables can be predicted, is tested with a 5-fold test, which has been repeated 20 times with different divisions. All of the other variables are fixed in this test. The results can be found in Figure 9.5.

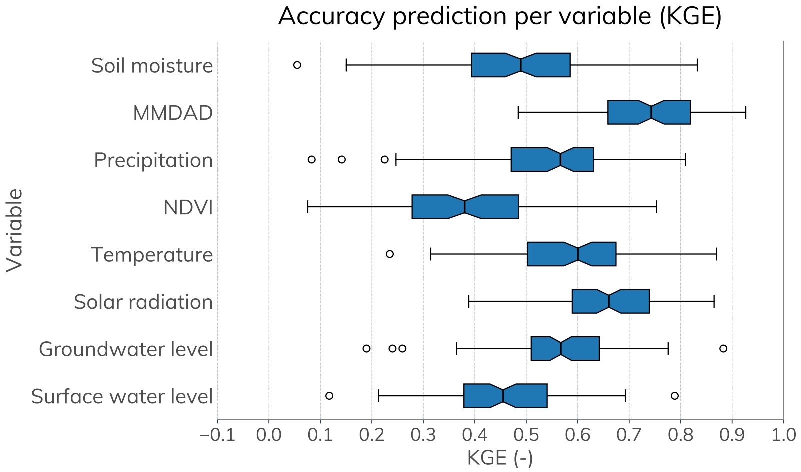


Figure 9.5: KGE per variable when all other variables are fixed. Tested 5-fold with 20 random repetitions per variable.

As can be expected, the MMDAD can be predicted best. The model is optimised for this variable and also has the most connections to other variables. The solar radiation is also relatively well predictable. A likely reason for this is the fact that the solar radiation variable has three connections in the model and has a high seasonal dependency, which makes a prediction in general relatively easy. The precipitation, which was able to predict the discharge the best of the variables (see Section 9.1.4), is averagely predictable. The NDVI is not so useful to predict the MMDAD, as well as difficult to predict itself. This makes NDVI in the prediction sense a lesser useful variable in the model.

9.1.6. Predicting all days

Same model, testing all days

In this thesis, the modelling has been focused on the monthly maximum discharge days. However, in practice, when an extreme event has to be predicted, it is likely that this day is unclear. Therefore, a number of tests has been conducted to determine the predictions when the data is tested on all days.

The first test is testing the model on all of the days. This means that the subset periods heavily overlap. The KGE that resulted from this test was 0.04. This is mainly due to the variance and mean parts of the KGE, which were 1.45 and 1.82 respectively. Especially the too high mean has a have a high influence on the KGE, as has been determined in Section 7.4.3. This poor performance is very logical, as the model's goal was to model monthly maximums. As only maximums were fed into the training set, the model is likely to predict high values as well.

Therefore, it is more relevant to look a the predictions of the monthly maximums. The KGE that came out of this test, was 0.72. So, the model still performs about equal when also other days are tested. However, only 16% of the dates on which this event happened was correct. This is still higher than the fully random 3%. The average distance between the days that the actual event was and the predicted, was 6.5. Furthermore, 61% of the predicted days were within 4 days of the actual event, whereas a random distance would be approximately 10 days. Thus, when the date of the event is predicted wrongly, the model still predicts a date within the same weather pattern. Lastly, in 62% of the cases, the model predicted the date too late in the month.

Model based on all days

It is also possible to set up the model itself for all days in the dataset. This completely ignores the conditional independence assumption, as all subset periods of data now highly overlap. The KGE that was acquired when testing this model was 0.73, when 5-fold tested with 20 random repetitions. This suggests that the conditional independence assumption is not needed for predictions for these kind of models. However, as approximately 30 times more data is used for this model, it can be expected that the model based on all days should perform better. As this is not the case, this might be an indication that there is a negative effect of the non-independent dataset.

9.2. Benchmark models

It is interesting to see how well other models would perform in this catchment. That is why, four different kind of models were selected to see how well they perform on the same target variable, the MMDAD, as well as (as much as possible) the same input values.

9.2.1. Saturated Bayesian network

Creating a model that has all the connections, has a number of advantages. First of all, it eliminates the need of having knowledge about the catchment in order to create a physically-based Bayesian network (BN) layout. Of course, this also has the disadvantage that the model is less intuitive and features inconsistent physical relations. Therefore, not all conditioning makes sense. Secondly, there is no need to use the recursive correlation equation (Equation (2.17)), which can be complicated to implement. In this thesis, Uninet is used to acquire these correlation coefficients. The correlation matrix R used for a saturated BN has the normal rank correlation $r_{r,\text{norm}}$ (Equation (2.12)) on all indices. Lastly, because this model uses all information available, the model has a marginally better accuracy in predicting the target variable MMDAD.

Using the same BN parameters as the unsaturated model, except for the layout, the average KGE became on average 0.74 per fold, compared to 0.73 for the unsaturated BN.

9.2.2. SOBEK model

SOBEK is a software suite that has been created by Deltares, which offers water system calculations that follow the Saint-Venants equations[1]. It is widely used in the Dutch water sector, and many water boards use it for multiple purposes, among which as their official peak flow modelling. This is done to check whether they comply with the agreement *Nationaal Bestuursakkoord Water* (National Governmental Agreement Water), which states several inundation limits (Rijksoverheid et al., 2003). Waterschap Drents Overijsselse Delta (WDODelta) also uses a SOBEK model for their water flows, for example, to check the effect of any alterations that they might implement in their system. Other SOBEK models are constructed for a multitude

[1] See https://www.deltares.nl/nl/software/sobek-suite/

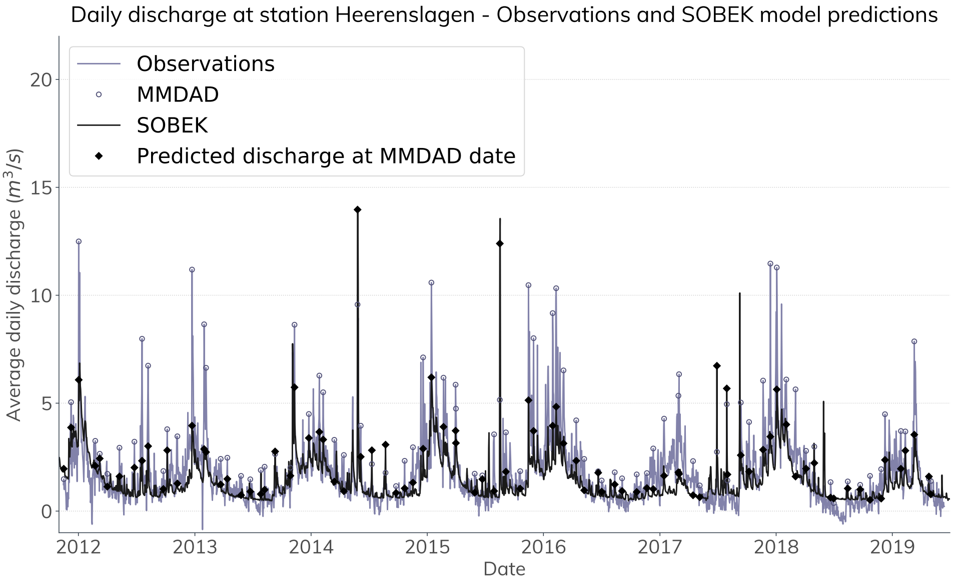


Figure 9.6: Daily average results for the SOBEK model as well as the discharge observations at the measurement station Heerenslagen. The values at the MMDAD events are marked as well.

of goals: among others monitoring effects on discharge and water level extremes, checking the water quality, water balance and predicting aridity. The model used in this thesis is made for predicting discharge peaks in a quick fashion, a so-called decision supporting system. It is maintained by the company Deltares, and changes are added regularly. The model calculates solutions in a numerical fashion, with timesteps in the order of 1 minute to 1 hour.

A slightly altered version of this model is used to benchmark the results against (no alterations were made in or closely surrounding the catchment). The same precipitation and potential evaporation values were taken as those that are used in Section 3.4. Wind and water temperature were not defined because these are not acquired for this research. The discharge is taken from the same stretch as where the main measurement station is situated. The parameters have been fitted beforehand, but despite this, the running of a model evaluation still takes 25 hours, when timesteps of 10 minutes are used over 8 years of data.

RESULTS

The general model results can be found in Figure 9.6. The model seems to be predicting a similar base flow as the measurements. However, the volatility seems to be far off and some peaks and periods of higher discharge, such as the winter of 2012-2013 are underestimated. The KGE for all dates was only 0.576, and the α factor of the volatility was 0.700, which is very low for such a model. Moreover, for the monthly maximums, the model performed even worse: it acquired only a KGE of 0.400, as all peaks except for two, were (sometimes highly) underestimated. The following reasons for this low value are:

- The model has been made for the whole network of the water board. Therefore, it is possible that the fitting of this section was made worse, because it had to comply with other measurements downstream of the catchment.

- The fitting of the discharge at the Steenwijker Aa did not deliver as good results as that of other catchments (De Graaf and Rusticus, 2013).
- There were only two months of data used in fitting the network (De Graaf and Rusticus, 2013). This can cause the model to overfit that event, thus producing worse fits in other events (for example in a different season).
- Weir level differences were not implemented in the model. To store water for dry periods, the water level is artificially raised in spring and summer. This parameter is implemented in the BN (surface water level, Section 3.3.6), but not in de SOBEK model.

However, this is the model that is used right now to predict high discharges at the water board. A BN provides a simple tool in which extreme events can be predicted in a computationally inexpensive and with a better perfomance than the SOBEK model.

9.2.3. NEURAL NETWORK

The data of model 1 has been used in an (artificial) neural network (NN). This is a model that is rooted on the working of the brain. Via simple switch-like functions in a great number of neurons, the model is able to make predictions of various phenomena. It is usually not traceable what the exact reasoning is behind the choices for the parameters, but for the target value, the method is capable of calculating a wide range of types of predictions. These types can be highly abstract, such as predicting an image either containing a cat or a dog. This means that the method is often called ’black box’: the reasoning behind the model’s ’choice’ cannot be explained easily. This can be unfavourable for governments because they often have to explain their reasoning.

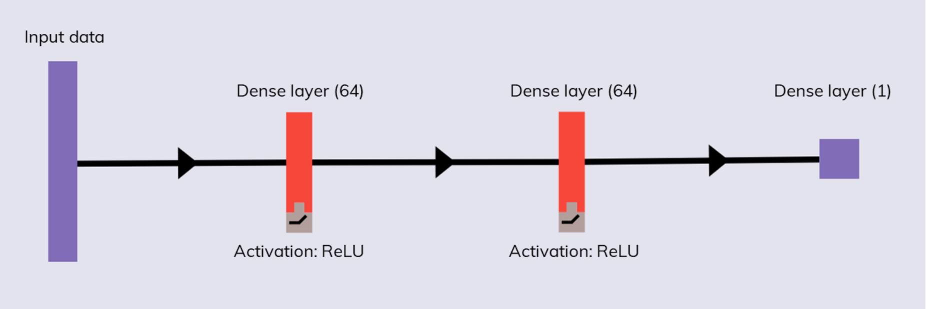


Figure 9.7: Layout of the Tensorflow artificial neural network that is used. It has two dense layers with 64 neurons and to get the target value, a single dense layer of 1 neuron.

The model that is used is shown in Figure 9.7 and consists of two hidden, deep layers of 64 neurons. This is the type of model that is advised by Google’s NN package Tensorflow, in a tutorial for similar regression models by Chollet (2017). The core of this model consists of Keras elements.

The average KGE for the model shown in Figure 9.7 was 0.58, and fitting and predicting once with the model took on average 18.4 seconds. A possible explanation for the low performance is the relatively small number of data points. NNs typically need a large training dataset.

9.2.4. MULTIPLE LINEAR REGRESSION

A basic method to predict a variable from another variable is linear regression ($y = \alpha x + \beta$). It is also possible to predict values V_{sim} for any number of variables with the formula

$$V_{\text{sim}} = X\mathbb{A}, \tag{9.2}$$

with X defined as

$$X = \begin{bmatrix} 1 & v_1^1 & \dots & v_1^d \\ \vdots & \vdots & \ddots & \vdots \\ 1 & v_n^1 & \dots & v_n^d \end{bmatrix},$$

where $v_1^1, \dots, v_m^n$ are the values of the first variable and so forth. The parameters $\mathbb{A} = \alpha_1, \dots, \alpha_n$ can be calculated with the normal equation:

$$\mathbb{A} = (X^T X)^{-1} X^T V_{\text{target}}, \tag{9.3}$$

where V_{target} are the target values ($m \times 1$). It is not possible to return a whole probability distribution for any of the prediction variables or the target variable (Ng, 2011). The 1s in combination with an additional α represent the offset of the origin, or symmetrically, the factor β in the bivariate linear regression formula. The method has been tested in Python with k-fold cross validation (see Section 4.3). In this case, the X in Equation (9.2) is different than the X from Equation (9.3). The average KGE was 0.70. The average time of a model evaluation takes 0.003 seconds.

9.3. COMPARISON MODELS

In Table 9.1, the different models discussed in this section can be compared. This table shows that, for the modelling goal of this thesis, the unsaturated Bayesian network is optimal. This offers advantages in modelling time, computational time and keeps physical relations. Moreover, other variables can also be predicted from fixed variables. Meanwhile, the methods keeps returning complete (conditioned) probability distributions. In comparison with the multiple linear regression, the unsaturated Bayesian network performs only slightly better in predictions. As the multiple linear regression is easier to set up, this might be an alternative in certain cases, where the other advantages of the unsatured Bayesian network are not of importance. The same holds true for a saturated Bayesian network, which results in even higher KGEs.

Method	Avg. KGE (-)	Avg. duration per run (s)	Probability distribution returned
Bayesian network	0.73	5	Yes
Saturated BN	0.74	5	Yes
SOBEK	0.40	45 hours	No
Neural network	0.58	18.4	No
Mult. linear regr.	0.70	0.03	No

Method	Conditioning	Black box	Preparation effort
Bayesian network	Physically valid relations	No	Low
Saturated BN	All, including impossible relations	No	Very low
SOBEK	Only spatially	No	High
Neural network	No	Yes	Low
Mult. linear regr.	No	No	Very low

Table 9.1: Comparison between different models tested in this research.

10

Discussion

In this chapter, the methodology, acquired results and determination of the optimal model will be put into context. This is done by looking at multiple topics: a general evaluation and interpretation of the research, potential limitations in the study, and a comparison with similar research. As the application of the method proposed in this thesis as a hydrologic model is still very novel, there are not many sources to compare the research against.

10.1. Model principles

In this research, the focus is completely on a single catchment and creating the best as possible model for the prediction of the monthly maximum daily average discharge (MMDAD) for this catchment. This is fundamentally different than for example Paprotny and Morales-Nápoles (2017), that use many catchments and strive to make a single model that can estimate the annual maximum runoff of catchments. The variables of this model are mainly catchment parameters such as the area and average slope of the catchment. At the other end, the work by Couasnon et al. (2018) focuses on very directly related variables: discharges predicted by the discharge upstream. This means that the case study proposed in this research is the first catchment in which a Bayesian network is made that is predominantly based on meteo-hydrological variables.

This means that the catchment variables have a different role in this thesis in contrast to Paprotny and Morales-Nápoles (2017). In this thesis, the workings of the meteo-hydrological variables are modelled by the Bayesian network (BN) nodes, which is expressed through its cumulative distribution function (CDF), whereas in Paprotny and Morales-Nápoles (2017), these are modelled by the arcs, and therefore the copulas. The opposite is true for the catchment parameters (i.e. slope, area etc.), which is inexplicitly modelled by the copulas in this thesis, whereas these are used as nodes in Paprotny and Morales-Nápoles (2017). Couasnon et al. (2018) takes a similar approach, although instead of the meteo-hydrological variables, it mostly uses discharges, which are more directly correlated to the downstream discharge than the meteorological variables.

This approach also means that the model is based on time-dependent variables. This has as a consequence that the variables are likely to being also dependent of itself, i.e. show autocorrelation. This means that an additional factor influences the variable, which cannot be modelled easily in the BN: the previous values of the same variable. As mentioned in Section 3.3.2, in this thesis, this phenomenon is minimised by using maximum monthly values (MMDAD). Couasnon et al. (2018) uses a way more frequent interval, using mean daily discharges. They assume that because of fact that the autocorrelation for their case drops rapidly, no issues can be expected from this. However, a lag of 1 still produces significant correlations in their research. It is interesting to see the influence of autocorrelation on the predictions in future research. Paprotny and Morales-Nápoles (2017), on the other hand, uses a spatial dataset and only one row per catchment, predicting the mean of the yearly maximum discharges. Influences of potential spatial correlation are not tackled in this paper, which is also an interesting subject for future research.

10.2. Case study

The goal of this research is to make a model of a single catchment, as a case study, such that several techniques and methods can be analysed in depth. This is in contrast with research by Paprotny and Morales-Nápoles (2017); Sanjaya (2018); Torres Alves (2018), which tested a multitude of catchments. These researches did not attempt to create an optimal discharge model per single catchment, but merely an overarching model for all these catchments.

In order to do research on a case study which is also useful for various other catchments, a very typical Dutch catchment has been chosen. This means that many Dutch catchments should be able to be comparable to this catchment, and conclusions made in this thesis are more likely also applicable to these other catchments. In order for this to be true, the chosen catchment should be representative of other lowland, partially managed, catchments.

However, no research has been put into verifying whether it is indeed representative. This is a vulnerability to the research, as it is possible that other catchments give different accuracies when tested on a similar model. However, the data used for this research is available for many catchments in The Netherlands and other regions, such that a performance test of the BN should not be difficult in other catchments.

Another reason for the selection of this catchment, is that the data availability was well. Moreover, an outline of the catchment boundary could be made such that the data could be verified with a water balance (see Section 3.4).

To exclude additional influences to the catchment, the case study could is preferred to not be predominantly managed. In this case, the discharge is too much dependent on the view of the manager, instead of the other variables.

10.3. Data

The data that has been used for this research is acquired from different sources, which not all have been scientifically quality-checked, in contrast to the data used in for example Paprotny and Morales-Nápoles (2017). Therefore, in this research, the data has been examined for its quality. As the model should and does still work in circumstances in which the measurements record a bias, the data is not meticulously tested. This is in contrast with highly fundamental research, in which the data needs to be highly accurate.

This means that it is possible that there are some errors in the measurements, which have gone unnoticed. To exclude errors as much as possible, different sources of data have been compared and the final dataset has been verified as much as possible in the Budyko framewerk Section 3.4. On top of that, in Section 7.4, a comprehensive analysis is done in what the effect of error could be, which is not large in the case of most of the variables.

During the timespan of which the data is used, a number of changes have been made to the river stream. Most of these changes represented a minor section of the catchment, but changes in the outflow pattern cannot be ruled out.

10.4. COPULA ASSUMPTION AND MULTIVARIATE NORMAL METHOD

In Chapter 5, the assumption of combined probability distributions of the variables following a Gaussian copula is tested. Especially noticeable is the fact that for a small majority of the variable pairs, the Gaussian copula does not satisfy the one-level bootstrapped test. This shows that the Gaussian copula is not a perfect description of the combined probability distributions, for 13 of the variable pairs. However, finding a single copula that can model all of the pairs perfectly is not the goal of this research.

What Chapter 2 does show, is that in general, the Frank copula would have been a better fit to variable pairs. The Frank copula is more complicated to implement than the multivariate normal method, and therefore not strictly a better (user-friendly) model than the Gaussian copula.

What is clear from Chapter 2, is that 11 variable pairs showed some kind of tail dependence, which Gaussian copulas fail to model. However, this leaves a majority of variable pairs in which the Gaussian copula does model the lack of tail dependence well. This also shows in the fact that the Frank copula, which also has no tail dependence, fits the best for many variable pairs.

All in all, the Gaussian copula turned out to be not strictly optimal, but still has a good basis to be used. Moreover, the goal was to make a good overall model, not making every element perfect.

One of the reasons for the usage of the multivariate Gaussian copula, is that it prevents need for using the highly complicated vine-copula structure. No similar hydrologic models as proposed in this thesis have been made using vine-copulas. There are other calculations done in hydrology that do use vine-copulas, such as Gräler et al. (2013). However, these use significantly less nodes, such that the number of vines, which can be calculated by Equation (B.3), is significantly less.

10.5. MARGINAL DISTRIBUTION

In Chapter 6, the choice has been made to opt for volatile marginal distributions which fit the empirical distributions very closely, and implement the same type of distribution for all variables. The Gaussian mixture model is selected in contrast to the empirical cumulative distribution function (ECDF) as this allows for extreme uniform values to be translated to values that have not been measured before. Paprotny and Morales-Nápoles (2017) use the ECDFs for all of the non-target variables. Their study included as many European catchments with measurements that suit these methods as possible. This means that their dataset provides a wide range of measurements, from 1841 stations. Therefore, it is likely that for any new catchments that are tested, the variables are included within the minimal and maximal value in the ECDF. In this thesis, it is likely that, as only about 8 years of data are used, it is likely that in the future new extremes will be measured.

The Gaussian mixture model, which is used, has been implemented before by Couasnon et al. (2018). In this paper, the rest of the variables is fitted with a generalised extreme value (GEV) distribution. Paprotny and Morales-Nápoles (2017) also use this distribution for their discharge measurements. This thesis proposes a shift function of the uniform variables, to make the model underestimate the extreme discharges less. This happens because the closely fitting marginal distributions predict very little values in its extremes. The marginals that use the GEV distribution does not have this problem, as it's CDF is smoother and approaches 1 relatively more gradually. The shift function still benefits from the use of a theoretical distribution, as the shift function only exaggerates what is happening in the fringes. When the input in these edges is highly unfavourable, the output of the shift function cannot be highly favourable behaviour. This boils down to two disadvantages: (1.) the step function, which is the ECDF, make that only a small number of steps is actually shifted and the shift is highly dependent on the location of the step and even more importantly (2.) as the ECDF has a fixed point at the highest value where $u = 1$, the complete top step cannot be shifted whatsoever.

10.6. Order of parameter optimisation

The optimisation of the BN has been done through consecutively finding the best-fitting parameters. However, this leaves out certain parameter combinations that have not been tested.

The reason for this lack of thoroughly testing all combinations is that the computational time would have been increased a lot if a 20 times randomly repeated 5-fold test would have been conducted for all the combinations. Especially parameters that have been tested on a large number of values, such as the number of samples, would make this method very slow. Moreover, testing each parameter consecutively allows for a more clear visualisation of the test results.

It is assumed that, because the differences between various parameters were low, it is not likely that an untested combination of variables performs significantly better than the final model does.

10.7. Testing optimisations

Except for the copula assumption, most of the testing has been conducted a posteriori with as goal variable the MMDAD. There is a number of reasons why this is done: firstly, in this thesis, the main target was predicting the MMDAD. The copulas modelling the multivariate distribution perfectly was also important, but were ranked secondary. Moreover, these kinds of optimisations are often more understandable for users as they often also want to predict such a target variable. Functions like the Akaike information criterion (AIC) are often regarded as less intuitive.

On this point, the work from Paprotny and Morales-Nápoles (2017) looks at this differently. For example, when fitting a marginal distribution for the discharge (the target variable in this paper), the AIC is used on the fit of the CDF to the data. This means, the optimal function is determined a priori.

The main test method used throughout this research is the Kling-Gupta efficiency (KGE). This coefficient gives an overall score based on the difference between the predicted outcome and the observations, based on three submetrics. However, how hard more extreme inaccuracies should be penalised in the score, is always up for debate. The KGE uses the Pearson correlation coefficient, but other methods (mean squared error for example) could also be possible, depending on the preferred relative penalty for high inaccuracies.

If the scores of this model should be compared to other research, the Nash-Sutcliffe efficiency (NSE) is a more common metric. The main reason for this, is that the NSE is an older metric. Therefore, it is likely that in the future, a higher share of research uses the KGE.

The KGE is, in most of the cases, used for the target variable MMDAD. As a goal of the model is to also being able to predict other variables as well, this might in some cases have given a different result. However, as the MMDAD is regarded as the most important variable to predict and a single answer was required, testing on this variable only was chosen in most of the cases.

These tests are almost all conducted with a k-fold test, where 5-folds were used. This was generally repeated 10 or 20 times with new random folds. The performance between using different parameters was compared with boxplots of these outcomes. Differences between these boxplots were often small. Therefore, no always a clear optimum could be selected. Repeating the k-fold test more often, would have probably not helped, as a repetition of the test of 50 to 100 times was already a lot.

10.8. Bayesian network layout

In this thesis, the configuration of the BN is made based on catchment processes in a single catchment. Every step is thoroughly analysed based the strategy that followed from the criteria (see Section 8.1) and if no decisive argument could be made about two contradicting local configurations, predictions within the model decided upon the implementation. Other research, such as Couasnon et al. (2018) is more spatially based, or has executed the configuration less descriptive, such as Paprotny and Morales-Nápoles (2017).

As the implementation of these criteria was often multi-interpretable. Moreover, not all tests gave significant differences in results. Therefore, the final model setup cannot be regarded as the overall best, for any user or any purpose. However, with this two-step method, anyone who repeats this test, will likely end up with the same model.

11 Conclusions and recommendations

For many years, improvements have been made in hydrologic modelling in catchments. For a long time, a vast amount of data has been collected, which contributes to the accuracy of these hydrologic models. Since the last decades, this has been complemented by various satellite measurements. This makes way for a new generation of models, which relies heavily on the availability of data over the last decades.

This removes the need for modelling based on the physical structure of the catchment. For conceptual models the area, storage and runoff parameters have to be determined for a catchment. In lumped and semi-distributed models, this is inherently flawed because of the heterogeneity of catchments. Moreover, these models can show equifinality, in which significantly different parameter sets perform equally optimal. These problems are not present many data driven methods, such as the non-parametric Bayesian network (NPBN). This statistical method is implemented in this thesis to describe hydrologic processes in a lowland catchment in The Netherlands. This thesis posed a twofold research objective.

(i) Firstly, the goal was to create a hydrologic model that is formed by a Bayesian network (BN) which is easily usable by managers and researchers of the catchment and delivers a high accuracy in predicting variables for the catchment of the Vledder, Wapserveense and Steenwijker Aa. More specifically, as the main target variable, the monthly maximum daily average discharge (MMDAD) has been selected.

(ii) Secondly, the goal of this thesis was to analyse the performance of the model that was constructed during the completion of objective (i) and benchmark this against other model types that predict the discharge in a catchment. Combined, the objectives form the research question:

> What is the optimal setup of a Bayesian network hydrologic model in a lowland catchment, and how does it perform?

The first objective has been addressed in several subproblems. First of all, the most favourable method to construct a NPBN was chosen in Chapter 2. This has not been an exhaustive research where all methods have been made into a model and the most suitable has been chosen. However, in Chapter 5, the method has been compared to other copulas and has been proved to be sufficiently fitting to the data.

There are various other methods to approach a BN. These differ also in to what degree they are related to the NPBN with the multivariate normal (MVN) approach. Future work can compare different approaches to a hydrologic model BN to the one proposed in this thesis.

Secondly, the variables to be used in a model have been determined in Chapter 3. This has been done on the basis of availability of data, presumed relation to other relevant variables and its actual correlation between themselves. This thesis provides an arrange of variables that meet the minimal frequency and time frame, and are highly relevant for users and discharge predictions. Of these variables, an initial model has been made for the tests that followed.

It is recommended that users of similar models look critically at their own case: the creation of such a model is dependent on the availability of good quality data sources in their catchment. It might be possible that not all variables that are used in the model of this thesis, are also available (with sufficient longevity, frequency and quality) in that catchment. The opposite is also possible: additional variables such as seepage, or other measurement locations in different subcatchments can also be added in models of other catchments.

In the future, other sources for the same type of data can be used when they have have acquired enough longevity. Examples of these are Soil Moisture Active Passive (SMAP) for soil moisture data and other groundwater level stations. Additionally, using the exact model for a longer time might result in higher accurate predictions. The data used in this thesis consists of 91 rows (timesteps). If the model is recreated in a number of years, this number increases significantly, which has the potential of increasing the accuracy of the model.

The data is aggregated per month because this of the this forms an optimum between on one hand, having variables that a certain degree of temporal independence, and having enough rows. If for example, a week was taken as the interval between variables, the number of rows increased to approximately 390, whereas the temporal independence is significantly reduced. The consequences of this can be research in future work.

The third step answering the first research objective was to create a cumulative distribution function (CDF) that fits the data well and lets the model perform well predicting the MMDAD. This has been covered in Chapter 6. Three different CDF functions have been proposed. The first is a strictly empirical one and is expressed as a step function based directly on the data. This means that this is the model with the least effort to make. However, unaltered, it is not possible to extrapolate with this function. The second is a novel function, based on the logistic function. The final model was the Gaussian mixture model, which was selected as the best function. The latter two models found very similar fitted functions as they had a relatively high variance. It is likely there is no CDF that is used for all variables, that performs significantly better than these the Gaussian mixture model. It was apparent Chapter 6 that CDFs with less variance resulted in worse Kling-Gupta efficiencys (KGEs). A very high variance CDF, with many parameters, did not improve the results compared to one with a relatively moderate number. Therefore, it can be argued that the optimal CDF has been found for similar cases as this one. However, the CDF has not been optimised per individual variable but the same optimal number of parameters and type is used on all variables. Additionally, the shift function that is produced in Section 6.6 is up to the user's preference. Therefore, the chosen implementation of the CDFs cannot unambiguously be regarded as the optimal implementation all purposes of this model.

As fourth step, three parameters of the model have been optimised in Chapter 7. This resulted in using the *mean* of *500* samples sampled from copulas that have been fitted by making variables of aggregated samples of *8.5* days before the MMDAD event. These were additional steps in answering the research question of creating an optimal model.

It is recommended to verify these parameters in similar models of other catchments. This can rule out whether these parameters are generally optimal for these kinds of models, or just for this specific case.

Lastly, the layout of the nodes and connections in the BN (see Appendix A) has been discussed in Chapter 8. This gave rise to a range of layouts that comply with the criteria posed in the same chapter, which all fulfilled the first objective of this thesis. A single optimal model had to be selected by testing the predictions that a certain layout could produce. This layout in combination with the parameters and settings determined before, makes up the model is the model that answers the first part of the research question (see Chapter 11).

It is recommended for creators of similar models to first make physical-based decisions for their specific BN, based on leading processes in their catchment. If ambiguities arise, they can also resort to performance-based choices.

Due to several subproblems mentioned above that have not been tested exhaustively and a number of ambiguities in the determination of the most favourable model, it cannot be unconditionally determined that the model proposed in this thesis, is in fact the optimal model for the catchment of the Vledder, Wapserveense and Steenwijker Aa. However, on all levels of the NPBN, this thesis has striven to create the best as possible model, both performance-wise, physical-wise and usability-wise. Therefore, in the effort-sense of the word 'optimal', an optimal NPBN model (that uses the MVN method) has indeed been constructed.

The second objective, the analysis of the model performance and benchmarking it against other models, has been addressed on multiple levels as well. The first level is the testing of the assumption that the Gaussian copula fits the variable pairs sufficiently in Chapter 5. This assumption has been thoroughly examined and it can be concluded that the Gaussian copula fits many variable pairs well, in spite it was not always the best fitting copula for all pairs. The methods used in this chapter have been very thorough and therefore it cannot be expected that using other tests gives significantly different results.

Indirectly related to this objective, is the analysis of how errors in the data would influence model performance. As it is not possible to determine the actual errors that were in the data, an approach has been taken that evaluates the sensibility to changes in the variables, which can be seen as 'adding new errors in to data'. These tests helped putting the performance of the model in relation to the data quality. It showed that changes to variables not directly connected with the MMDAD did not have a significant influence on predictions of the MM-

DAD. Moreover, the influence of artificial random errors was very low. The influence of bias was higher and also depended on the sign of the correlation: negatively correlated variables showed actually improvement in predicting the MMDAD in cases with low underestimations.

It is recommended, if a more precise prediction is expected, to quantify actual errors in the data. This can be done by in situ measurements of the same variable. An example of this are the discharge measurements with several side looking doppler measurements devices, which can be verified by executing more acoustic doppler current profiler (ADCP) measurements to verify or improve regular measurements.

Thirdly, the KGE, a metric that determines the accuracy a model, has been calculated for the final model in Section 8.5. This indicated that the model performed reasonably. As an additional check, a complete k-fold cross validation has been performed, over all fitted parts of the model. A k-fold test divides the data into k folds and fits the data using k-1 folds and tests the last fold. This test It gave a similar result to other (randomly repeated k-fold) tests. The k-fold test can be regarded as one of the definitive answer to the second part of the research question. To give a more in-depth review of the model performance, a number other analyses have been performed on the prediction of the model. The relationship between the prediction error and the observed MMDAD has been calculated in this section as well. This makes it possible to analyse what discharge magnitudes likely give raise to what error magnitudes and whether this is an under- or overestimation. The research objective of analysing the model performance has become more detailed with Figure 9.3. In the same section, two more analyses have been done: the prediction of the target variable MMDAD when not all other variables are fixed, and the prediction of other variables than the target variable. This gives a broader overview of the general model performance, instead of just looking at the primary function of the model.

Lastly, four different models have been introduced to benchmark the model performance. The reason for this, is that a model performance is assess to regard on its own. When compared to other models, the context in relation to what is possible with this data, becomes more clear. The production of these other methods, has not been an exhaustive effort, such as has been done with the BN in this research. Therefore, the result that the BN performs better and similarly to other models, cannot be taken as a definitive result. However, it does give a more contextual indication how well the model is performing.

It is recommended for Waterschap Drents Overijsselse Delta (WDODelta) to improve their SOBEK model if they want to keep using this for modelling the discharge at station Heerenslagen, as the model performs dissatisfactory now. Possible reasons for this are that the model is overarching multiple catchment and it is fitted better to the other catchments. Moreover, it is likely that it is better fitted to baseflow than peaks, although these are often more important for floods and pump capacities. Another recommendation is to benchmark the model proposed in this thesis against other models (such as the neural network (NN)) if they are optimised in a similar fashion as in this research. This will give a fairer comparison between the (near) optimal performance of models.

Based on these conclusions, managers of the catchment of the Vledder, Wapserveense and Steenwijker Aa could use the model proposed in this thesis. The model has proven to predict the target variable, the MMDAD, decently and has other benefits, such as that it also provides a probability distribution of the predicted variable(s) and is able to predict when not all other variables are known. This might happen when a measurement station stops functioning or when reliable predictions can only be made for a number of variables, if the discharge needs to be predicted further in the future than a single day. The method is ready to be used through a combination of the commercially available software Uninet[1], in combination with the Python package `copulabayesnet`[2]. It is recommended that a Equation (2.17) is added to `copulabayesnet`, or that the prediction and testing capabilities of `copulabayesnet` are added to Uninet, such that everything can be run from a single platform. The rudimentary application this thesis proposes, is the prediction of the monthly discharge peak. However, this thesis does not make a detailed description about the precise practical applications of the model. Examples could be, flood protection, pump capacity prediction, determining what catchment processes influence high discharges to what extend. Therefore, the first approach for authors and operators of this model and similar models in other catchments, could be determining its professional application or applications.

In order to create a similar model in a different catchment, first of all, the recommendations mentioned above could be helpful to create and verify such a model. These are: checking the data availability (and potentially also quality), finding the optimal parameters, finding out the core catchment processes for the best BN layout.

Additionally, using models that have very similar settings to the model proposed in this thesis, can solve some interesting research questions. First of all, it can further verify the method proposed in this thesis. If the model also performs similarly in a significant number of other catchments, it can be believed that the method functions well in general, not just in the Vledder, Wapserveense and Steenwijk Aa catchment.

It is helpful to know the influence of the amount of data. For a BN model, the catchment used had a low amount of data with only eight years of data in the dataset, which was resampled to monthly data. It is interesting to see how similar catchments with a lot more data - or even - less data perform.

Testing other catchments could also be a method to see the influence of the part of the catchment that is managed. As this is a statistical method, the assumption is that variables (that is, the monthly aggregated values) behave randomly only affected by other variables, only being influenced by other variables. This is not the case for managed systems, as a person or system 'decides' upon a measure. When the share of water that is managed is low enough, the catchment as a whole can still be regarded sufficiently random. However, to find the maximum share of managed area, more catchments with different portions of management should be tested. Some management of water flows, such as the water level, happens automatically because a certain condition. This automation can in a way be regarded as only influenced by other variables that are already included in the BN, and therefore quasi random. Therefore, it would be interesting to see if it matters whether management happens automatically or manually.

In addition to the recommendations and potential for future work already given, there are

[1] Available on `https://lighttwist-software.com/uninet/`

[2] See Appendix G.4.

some research subjects not directly in line with this research, that are also very interesting. First of all, instead of predicting the MMDAD, a range of other target variables can also be used. Examples of this are different values in the set, such as the monthly minimum discharge or monthly average discharge, and different timeframes, such as weekly or yearly maximums, or a combination of these.

Moreover, research can be done in the creation of models that centre around other target variables, such as water level, or that try to maximise the prediction of all variables.

So all in all, in this research, a hydrologic model Bayesian network has been created for the catchment of the Vledder, Wapserveense and Steenwijker Aa, which has been optimised on several levels to predict the monthly maximum daily average discharge out of the catchment, and on other levels has been suited to fit the implementation and usage best. The accuracy of this model is decent and outperforms several other models. Although the method is readily applicable, there are some parts that can still be researched further.

References

Aas, K., Czado, C., Frigessi, A., and Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, 44(2):182–198, ISSN: 0167-6687, DOI: 10.1016/J.INSMATHECO.2007.02.001.

Actueel Hoogtebestand Nederland (2019). AHN3. https://www.pdok.nl/introductie/-/article/actueel-hoogtebestand-nederland-ahn3- (visited on 2019-11-12).

AdviesCommissie Schade Grondwater (2015). Schadeonderzoek Grondwateronttrekking Terwisscha. Technical report, AdviesCommissie Schade Grondwater, Utrecht.

Arbenz, P. (2013). Bayesian Copulae Distributions, with Application to Operational Risk Management-Some Comments. *Methodology and Computing in Applied Probability*, 15(1):105–108, ISSN: 15737713, DOI: 10.1007/s11009-011-9224-0.

Arora, V. K. (2002). The use of the aridity index to assess climate change effect on annual runoff. *Journal of Hydrology*, ISSN: 00221694, DOI: 10.1016/S0022-1694(02)00101-4.

Bedford, T. and Cooke, R. M. (2001). Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial Intelligence*, 32(1-4):245–268, ISSN: 10122443, DOI: 10.1023/A:1016725902970.

Bergström, S. (1976). Development and application of a conceptual runoff model for Scandinavian catchments. In *SMHI Report RHO 7*, page 134 pp., Norrköping.

Branch, M. A., Coleman, T. F., and Li, Y. (1999). Subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems. *SIAM Journal of Scientific Computing*, 21(1):1–23, ISSN: 10648275, DOI: 10.1137/S1064827595289108.

Budyko, M. I. and Miller, D. H. (1974). Climate and Life. *Academic press New York*, 508.

Cambridge English Dictionary (2020). MACHINE LEARNING | meaning in the Cambridge English Dictionary. https://dictionary.cambridge.org/dictionary/english/machine-learning (visited on 2020-2-11).

Chollet, F. (2017). Basic regression: Predict fuel efficiency (Tensorflow Tutorial). https://www.tensorflow.org/tutorials/keras/regression (visited on 2020-04-03).

Cooke, R. M., Kurowicka, D., Hanea, A. M., Morales, O., Ababei, D. A., Ale, B., and Roelen, A. (2007). Continuous/Discrete Non Parametric Bayesian Belief Nets with UNICORN and UNINET. http://resolver.tudelft.nl/uuid:ec6eb1df-2bb6-4aa0-98b2-e1990c03e4b0.

Couasnon, A. (2017). Characterizing flood hazard at two spatial scales with the use of stochastic models: An application to the contiguous United States of America and the Houston Ship Channel. Master's Thesis. https://repository.tudelft.nl/islandora/object/uuid%3Af90f1b6c-d5fa-4891-b287-e03c9fec4118?collection=education.

Couasnon, A., Sebastian, A., and Morales-Nápoles, O. (2018). A Copula-Based Bayesian Network for Modeling Compound Flood Hazard from Riverine and Coastal Interactions at the Catchment Scale: An Application to the Houston Ship Channel, Texas. *Water (Switzerland)*, 10(9), ISSN: 20734441, DOI: 10.3390/w10091190.

De Graaf, J. and Rusticus, R. (2013). Waterschap Reest en Wieden - Herijking WB21 wateropgave - Eindrapport. Technical report, Waterschap Reest en Wieden, Meppel.

Didan, K. (2015). MOD13A1 MODIS/Terra Vegetation Indices 16-Day L3 Global 500m SIN Grid V006 [dataset]. NASA EOSDIS Land Processes DAAC. https://doi.org/10.5067/MODIS/MOD13A1.006.

Didan, K., Barreto Munoz, A., Solano, R., and Huete, A. (2015). MODIS Vegetation Index User's Guide (MOD13 Series). Technical report, NASA, https://modis.gsfc.nasa.gov/data/dataprod/mod13.php.

Earth Observing System (2013). MODIS Satellite Sensor: bands and specifications. https://eos.com/modis-mcd43a4/ (visited on 2020-02-17).

Eaton, M. L. (1983). *Multivariate Statistics: a Vector Space Approach*. John Wiley and Sons, ISBN: 978-0-471-02776-8.

Embrechts, P., Lindskog, F., and Mcneil, E. (2001). Modelling dependence with copulas and applications to risk management. *Handbook of Heavy Tailed Distributions in Finance*, 8, DOI: 10.1016/B978-044450896-6.50010-8.

European Space Agency (2017). SMOS Data Products. http://bec.icm.csic.es/land-datasets.

European Space Agency (2020). Soil Moisture and Ocean Salinity (SMOS) - Soil Moisture [dataset]. https://earth.esa.int/web/guest/missions/esa-operational-eo-missions/smos.

Favre, A. C., Adlouni, S. E., Perreault, L., Thiémonge, N., and Bobée, B. (2004). Multivariate hydrological frequency analysis using copulas. *Water Resources Research*, 40(1), ISSN: 00431397, DOI: 10.1029/2003WR002456.

Friese Koerier (1953). Kanalisatiewerken in Vledder en Wapse vragen millioenen. Heerenveen, 1953-12-22. https://resolver.kb.nl/resolve?urn=ddd:010734895:mpeg21:a0132.

Friese Koerier (1965). Vledder: Geen Wateroverlast. Heerenveen, 1965-9-8. https://resolver.kb.nl/resolve?urn=ddd:010690031:mpeg21:a0146.

Genest, C. and Remillard, B. (2008). Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models. *Annales de l'institut Henri Poincare (B) Probability and Statistics*, 44(6):1096–1127, ISSN: 02460203, DOI: 10.1214/07-AIHP148.

Gentle, J. E. (2009). *Computational statistics*, volume 308. Springer.

González-Zamora, A., Sánchez, N., Gumuzzio, A., Piles, M., Olmedo, E., and Martínez-Fernández, J. (2015). Validation of SMOS L2 and L3 soil moisture products over the Duero basin at different spatial scales. In *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, volume 40, pages 1183–1188. International Society for Photogrammetry and Remote Sensing, ISSN: 16821750, DOI: 10.5194/isprsarchives-XL-7-W3-1183-2015.

Google Trends (2020). machine learning, statistical - Compare. https://trends.google.com/trends/explore?date=all&q=machine%20learning,statistical (visited on 2020-03-03).

Gräler, B., Van den Berg, M., Vandenberghe, S., Petroselli, A., Grimaldi, S., De Baets, B., and Verhoest, N. (2013). Multivariate return periods in hydrology: a critical and practical review focusing on synthetic design hydrograph estimation. *Hydrology and Earth System Sciences*, 17(4):1281–1296.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 377(1-2):80–91, ISSN: 0022-1694, DOI: 10.1016/J.JHYDROL.2009.08.003.

Hanea, A. and Harrington, W. (2009). Ordinal data mining for fine particles with non parametric continuous bayesian belief nets. *Information Processes Journal*, 4:280–286, https://www.researchgate.net/publication/228399374_Ordinal_Data_Mining_for_Fine_Particles_with_Non_Parametric_Continuous_Bayesian_Belief_Nets.

Hanea, A., Morales Napoles, O., and Ababei, D. (2015). Non-parametric Bayesian networks: Improving theory and reviewing applications. In *Reliability Engineering and System Safety*, volume 144, pages 265–284. Elsevier Ltd, ISSN: 09518320, DOI: 10.1016/j.ress.2015.07.027.

Hanea, A. M., Kurowicka, D., and Cooke, R. M. (2006). Hybrid method for quantifying and analyzing bayesian belief nets. In *Quality and Reliability Engineering International*, volume 22, pages 709–729. ISSN: 07488017, DOI: 10.1002/qre.808.

Helwig, N. E. (2017). Introduction to Normal Distribution. Technical report, University of Minnesota.

Hofert, M. (2008). Sampling Archimedean copulas. *Computational Statistics & Data Analysis*, 52(12):5163–5174, DOI: 10.1016/j.csda.2008.05.019.

Hoge Raad voor de Adel (1959). Register Hoge Raad voor de Adel. Available on: http://www.therightproductions.nl/hogeraadvanadel/index.php?id=109&wapen=46.

IPCC (2014). *Climate change 2014: synthesis report*. IPCC, ISBN: 9789291691432.

Joe, H. (2015). *Dependence Modeling with Copulas*. Chapman & Hall/CRC, Boca Raton, FL, ebook - pdf edition, ISBN: 978-1-4665-8323-8.

Kadaster (2019). Basisregistratie Topografie (BRT) TOPNL [dataset]. https://www.pdok.nl/downloads/-/article/basisregistratie-topografie-brt-topnl (visited on 2019-12-10).

Kelly, K. and Krzysztofowicz, R. (1997). A bivariate meta-Gaussian density for use in hydrology. *Stochastic Hydrol Hydraul*, 11:17–31, DOI: https://doi.org/10.1007/BF02428423.

Knoben, W. J. M., Freer, J. E., and Woods, R. A. (2019). Technical note: Inherent benchmark or not? Comparing NashSutcliffe and KlingGupta efficiency scores. *Hydrology and Earth System Sciences*, 23(10):4323–4331, ISSN: 1607-7938, DOI: 10.5194/hess-23-4323-2019.

Krzysztofowicz, R. (2002). Bayesian system for probabilistic river stage forecasting. *Journal of Hydrology*, (268):16–40, DOI: 10.1016/S0022-1694(02)00106-3.

Kurowicka, D. and Cooke, R. M. (2005). Distribution-Free Continuous Bayesian Belief Nets. In *Modern Statistical and Mathematical Methods in Reliability*, pages 309–322. World Scientific Publishing, London, tenth edition, ISBN: 981-256-356-3, DOI: 10.1142/9789812703378_0022.

Kurowicka, D. and Cooke, R. M. (2007). Sampling algorithms for generating joint uniform distributions using the vine-copula method. *Computational Statistics and Data Analysis*, 51(6):2889–2906, ISSN: 01679473, DOI: 10.1016/j.csda.2006.11.043.

Langendijk, D., De Vries, D., Fagel, M., Jansen, M., Reikens, B., and Nijboer, R. (2014). Op weg naar schoon water; Achtergronddocument Kaderrichtlijn Water 2e planperiode 2016-2021. Technical report, Waterschap Reest en Wieden.

Leeuwarder Courant (2016). Vitens halveert waterwinning in Drents-Friese Wold - Friesland - LC.nl. https://www.lc.nl/friesland/Vitens-halveert-waterwinning-in-Drents-Friese-Wold-21764326.html (visited on 2019-12-6).

Luxemburg, W. and Coenders, A. (2017). *CIE4440 Hydrological Processes and Measurements, Lecture Notes*. Delft.

Ministerie van Economische Zaken (2019). Basisregistratie Gewaspercelen (BRP) [dataset]. https://www.pdok.nl/introductie/-/article/basisregistratie-gewaspercelen-brp- (visited on 2019-12-10).

Molina, M., Fuentetaja, R., and Garrote, L. (2005). Hydrologic Models for Emergency Decision Support Using Bayesian Networks. pages 88–99. Springer, Berlin, Heidelberg, DOI: 10.1007/11518655_9.

Morales Nápoles, O. (2019). Bivariate Dependence (Copulas) - Lecture slides. TU Delft.

NASA (2000). Normalized Difference Vegetation Index (NDVI). https://earthobservatory.nasa.gov/features/MeasuringVegetation/measuring_vegetation_2.php (visited on 2020-02-17).

Nash, J. and Sutcliffe, J. (1970). River flow forecasting through conceptual models part I: A discussion of principles. *Journal of Hydrology*, 10(3):282–290, ISSN: 00221694, DOI: 10.1016/0022-1694(70)90255-6.

Nasr, A., Alfonso, L., and Morales Nápoles, O. (2018). Bayesian Networks and Data Driven Models for Estimating Extreme River Discharges Case Study: Magdalena-Cauca Basin, Colombia. *20th EGU General Assembly, EGU2018, Proceedings from the conference held 4-13 April, 2018 in Vienna, Austria, p.19813*, 20:19813.

Nelsen, R. (2006). *An Introduction to Copulas.* Springer, second edition, ISBN: 0-387-28659-4.

Ng, A. (2011). Linear Regression with multiple variables: Normal equation. Lecture Slides: Machine Learning. Coursera.

Nieuwsblad van het Noorden (1950). Overlast Vledder en Wapserveense Aa. Groningen, 1950-1-4. Available on: https://resolver.kb.nl/resolve?urn=ddd:010886514:mpeg21:a0112.

O'Dea, S. (2020). Information created globally 2010-2025. *Statista*. https://www.statista.com/statistics/871513/worldwide-data-created/ (visited on 2020-2-11).

Paprotny, D. (2017). Estimating extreme river discharges in Europe through a Bayesian network, Appendix: Diagnosis of underlying assumptions regarding the Non-Parametric Bayesian Networks. *Hydrology and Earth System Sciences*, 21(6):2615–2636, ISSN: 1607-7938, DOI: 10.5194/hess-21-2615-2017. https://www.hydrol-earth-syst-sci.net/21/2615/2017/.

Paprotny, D. and Morales-Nápoles, O. (2017). Estimating extreme river discharges in Europe through a Bayesian network. *Hydrology and Earth System Sciences*, 21(6):2615–2636, ISSN: 1607-7938, DOI: 10.5194/hess-21-2615-2017. https://www.hydrol-earth-syst-sci.net/21/2615/2017/.

Pearl, J. (1985). Bayesian Networks: a Model of Self-Activated Memory for Evidential Reasoning. In *Seventh Annual Conference of the Cognitive Science Society*, pages 1–20, Irvine, CA. Cognitive Science Society.

PIK (2020). Global radiation PIK Research Portal. https://www.pik-potsdam.de/services/climate-weather-potsdam/climate-diagrams/global-radiation (visited on 2020-04-08).

Renard, B. and Lang, M. (2007). Use of a Gaussian copula for multivariate extreme value analysis: some case studies in hydrology. *Advances in Water Resources*, 30:897 – 912, DOI: 10.1016/j.advwatres.2006.08.001.

Rijksoverheid, Interprovinciaal Overleg, Unie van Waterschappen, and Vereniging van Nederlandse Gemeenten (2003). Het Nationaal Bestuursakkoord Water. https://www.helpdeskwater.nl/onderwerpen/wetgeving-beleid/@176067/nationaal/.

RTV Drenthe (2014). Vledder 'moordstuw' verwijderd. 2014-2-17. https://www.rtvdrenthe.nl/nieuws/81971/Vledder-moordstuw-verwijderd# (visited on 2020-01-24).

Salvadori, G. and De Michele, C. (2002). 2-copulas In Statistical Hydrology: Theoretical Models of Bivariate Dependence. https://ui.adsabs.harvard.edu/abs/2002EGSGA..27.4330S.

Salvadori, G. and De Michele, C. (2004). Frequency analysis via copulas: Theoretical aspects and applications to hydrological events. *Water Resources Research*, 40(12), ISSN: 00431397, DOI: 10.1029/2004WR003133.

Salvadori, G. and De Michele, C. (2007). On the Use of Copulas in Hydrology: Theory and Practice. *J. Hydrol. Eng.*, 12(4):369–380, DOI: 10.1061/ASCE1084-0699200712:4369.

Sanjaya, S. (2018). The Application of Bayesian Network Model: Quantifying Riverine Flood Hazard in the Java Island. Additional thesis. Available on: https://repository.tudelft.nl/islandora/object/uuid%3Ae1beefe5-d403-4500-a461-fd373a599a8c?collection=education.

SAS Institute (2017). Canonical Maximum Likelihood Estimation (CMLE). https://documentation.sas.com/?docsetId=etsug&docsetTarget=etsug_copula_details21.htm&docsetVersion=14.3&locale=en (visited on 2020-04-12).

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:Publ. Inst. Statist. Univ. Paris.

The Editors of Encyclopaedia Britannica (2019). Student's t-test. https://www.britannica.com/science/Students-t-test (visited on 2020-05-11). Encyclopaedia Britannica, inc.

Torres Alves, G. (2018). Application of Bayesian Networks to Estimate River Discharges in Ecuador. Additional thesis. Available on: https://repository.tudelft.nl/islandora/object/uuid%3A6579ad40-1339-44b6-b9d4-6cae90190819?collection=education.

Wageningen UR (2019). Bodemkaart Nederland [dataset]. https://www.pdok.nl/introductie/-/article/bodemkaart-1-50-000 (visited on 2019-12-20).

Wang, W. and Wells, M. T. (2000). Model selection and semiparametric inference for bivariate failure-time data. *Journal of the American Statistical Association*, 95(449):62–72, DOI: 10.1080/01621459.2000.10473899.

Waterschap Drents Overijsselse Delta (2019). Peilgebieden WDODelta [dataset]. https://waterschap-wdodelta.opendata.arcgis.com/datasets/peilgebieden.

Yang H., Abbaspour, K. C., and Zhang Y. L. (2002). Desertification Control and Sandstorm Mitigation in the Area Encircling Beijing-with a Discussion on the Application of Bayesian Network and Hydrological Modeling. In *12th ISCO Conference*, Beijing. https://www.dora.lib4ri.ch/eawag/islandora/object/eawag:12063.

Zandstra, B. (2016). Gebiedsbeschrijving Vledder en Wapserveense Aa. Technical report, WDODelta, https://www.wdodelta.nl/mgd/files/waterbeheerplan_2016-2021_wdo_delta_ab_4_januari_26nov15.pdf.

Acronyms

ADCP acoustic doppler current profiler 29, 100

AIC Akaike information criterion 94

BN Bayesian network xi, xii, xiii, 1, 2, 3, 4, 7, 8, 14, 15, 16, 21, 23, 27, 30, 32, 33, 34, 35, 39, 40, 43, 44, 47, 51, 53, 57, 63, 67, 70, 73, 74, 76, 77, 78, 79, 86, 88, 90, 91, 92, 94, 95, 97, 98, 99, 100, 101, 117, 118, 119, 120

CDF cumulative distribution function viii, ix, xii, 5, 8, 9, 10, 11, 12, 14, 15, 17, 41, 44, 46, 47, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 63, 64, 65, 81, 82, 91, 94, 98, 99, 113, 114, 115

DAG directed, acyclic graph xi, 7, 74

ECDF empirical cumulative distribution function xii, 10, 44, 56, 57, 60, 93, 94, 129

ESA European Space Agency 33

GEV generalised extreme value 94

IID independent and identically distributed 26, 28, 32, 43

KGE Kling-Gupta efficiency ix, xi, xii, 41, 51, 53, 54, 55, 56, 60, 63, 64, 65, 68, 69, 70, 71, 77, 78, 81, 82, 83, 84, 85, 86, 87, 89, 94, 95, 98, 100

KNMI Koninklijk Nederlands Meteorologisch Instituut 29, 30, 31, 33, 35, 36, 38

MMDAD monthly maximum daily average discharge xi, xii, xiii, 3, 5, 10, 17, 27, 28, 35, 40, 43, 44, 45, 51, 57, 60, 61, 63, 66, 67, 68, 69, 76, 77, 78, 81, 83, 84, 85, 86, 87, 91, 92, 94, 95, 97, 98, 99, 100, 101, 102, 118

MODIS MODerate resolution Imaging Spectroradiometer 34, 35

MVN multivariate normal vii, 15, 16, 43, 98, 99

NASA National Aeronautics and Space Administration 34

NDVI normalised difference vegetation index viii, 34

NN neural network 2, 88, 89, 100

NPBN non-parametric Bayesian network xi, 1, 2, 3, 8, 14, 23, 73, 97, 98, 99

NSE Nash-Sutcliffe efficiency viii, 40, 41, 83, 95

PDF probability density function 9, 10, 11, 12, 13, 17, 18, 19, 53, 54, 113, 114

SMAP Soil Moisture Active Passive 98

SMOS Soil Moisture and Ocean Salinity 33, 34

WDODelta Waterschap Drents Overijsselse Delta 5, 22, 26, 27, 30, 31, 32, 36, 37, 38, 86, 100

Symbols

$\mathbf{1}$ Indicator function. 46, 52

A Conditioning (fixed) standard normal values. 16, 67

$\boldsymbol{\alpha}$ Volatility factor. 41

α cumulative distribution function (CDF) parameter. 52, 53, 54, 55, 56, 89

$\mathbb{A}$ Parameters of multiple linear regression. 89

b Width of a weir. 27

B $\max(V) - \min(V)$ 53, 55, 68, 69

β Bias factor. 41

c Copula function in its probability density function (PDF) form. 8, 9, 19

C Copula function in its CDF form. 8, 9, 45, 47

C_{emp} Empirical copula. 45, 46, 47

c_m Fitted friction factor weir formula. 27

$\bar{\mu}$ Conditional mean of the normal distribution. 16, 18

$\bar{r}$ Conditional correlation coefficient. 19

$\bar{R}$ Conditional correlation matrix of Gaussian copulas. 16, 18

C_θ^{Cl} Clayton copula in its CDF form. 11

C_θ^{Fr} Frank copula in its CDF form. 12

C_R^{Ga} Gaussian copula in its CDF form. 9

C_θ^{GH} Gumbel-Hougaard copula in its CDF form. 11

C_θ^{J} Joe copula in its CDF form. 13

cov Covariance. 14

d Number of variables. 8, 9, 13, 15, 16, 17, 89

$\mathbf{E}$ Expected value. 19

E_a Actual evaporation (i.e. evaporation and transpiration). 37, 38

E_p Potential evaporation (i.e. evaporation and transpiration). 37, 38

ϵ Error in the data. 67, 69, 70

f PDF of a variable. 8, 19

F Cumulative density function ~ converts values to uniform values. 8, 14, 15, 19, 41, 56, 114, 116

$\mathbb{F}$ Generic water flux. 37, 38

F_{emp} Empirical CDF. 52

f_{gm} PDF of the Gaussian mixture model. 54

F_{gm} CDF of the Gaussian mixture model. 54

F_{logi} CDF of the altered logistic function. 52, 53

$\mathbf{g}$ Scale factor shift function 58, 59

h Water level. 27, 32

H_0 Null hypothesis. 31

$\mathbf{h}$ Offset factor shift function 59, 60, 116

I Identity matrix. 9

Φ^{-1} Inverse CDF of the standard normal distribution. 9, 14, 15, 16

k Number of folds in k-fold testing. ix, xii, 41, 53, 56, 57, 63, 64, 67, 81, 82, 89, 100

K Number of parameters in $F(\cdot)$ fit function. 52, 53, 54, 55

l Lag. 44

L Lower triangular matrix of the Cholesky decomposition. 18

m Number of values in the dataset/variable. 13

μ Mean of the normal distribution. 16, 55, 114

μ_{act} Actual mean. 70

$\hat{\mu}_e$ μ subject to error. 67

μ_{obs} Mean of the observation values. 41, 70

μ_{sim} Mean of the simulation values. 41

n Number of samples. 18, 44, 46, 52, 89

$\mathscr{N}$ Normal distribution 16

N Repetition number. 46

NIR Near-infrared radiation. 34

$\hat{\vartheta}$ Copula parameter with the optimal fit. 13

p p-value. 46

P Precipitation. 37, 38

$\mathbb{P}$ Odd polynomial 52

ϕ PDF of the standard normal distribution. 19

Φ CDF of the standard normal distribution. 19

Φ_R Multivariate CDF of the standard normal distribution. 9, 15

Q Discharge. 27, 37, 40

r Generic correlation coefficient. 16

R Correlation matrix of Gaussian copulas. 9, 13, 14, 16, 18, 41, 67, 81, 82, 86

rg Rank vector of a variable. 14

ρ_{auto} (Pearson) autocorrelation. 44

$r_{\text{s,norm}}$ Normal rank correlation coefficient. 14, 16

ρ_{p} Pearson's correlation coefficient. 14, 41, 44, 45

r_{s} Spearman's rank correlation coefficient. 14

$\mathbb{R}$ Rank vector. 46

s Standard normal value 15, 16, 18, 19

S Vector with standard normal values 15, 16, 18

$\mathfrak{S}_{RMSD}$ Altered Cramér-von Mises statistic. 46, 47

$\mathfrak{S}_n$ Cramér-von Mises statistic. 45, 46

σ Standard deviation. 14, 41, 55, 68, 69

t Time/event. 40

t_e Last event in the data. 40

θ Archimedean copula parameter. 10, 11, 12, 13, 15

Θ Set of all (tested) valid copula parameters. 13

u Single value of a variable converted to a uniform value through $F(v)$. 8, 11, 12, 13, 15, 18, 19, 45, 46, 47, 58, 59, 94

U A set of uniform values (one value per variable). 15, 16

$u_{sd,\mathbf{h}}$ Shifted uniform value. 59

U' Vector of the values converted to a uniform values through $F(v)$, of a variable. 14

v Single value of a variable. 8, 15, 18, 19, 41, 44, 52, 53, 89

V A set of values (one value per variable). 15, 16, 89

ϑ Generic copula parameter. 13

$\mathcal{U}$ Uniform distribution. 18

VIS Visible light. 34

V' Vector of the values of a single variable. 14, 16

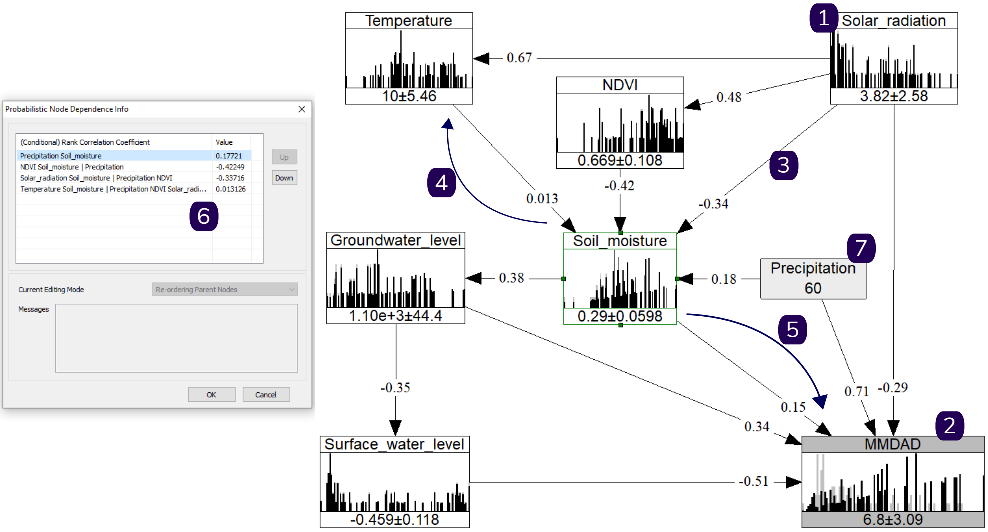
w Independent random standard normal value. 18

W Independent random standard normal vector. 18

z Number of conditioning (fixed) variables. 16, 17, 18

A

Anatomy of a Bayesian network



Probabilistic Node Dependence Info

(Conditional) Rank Correlation Coefficient	Value
Precipitation Soil_moisture	0.17721
NDVI Soil_moisture | Precipitation	-0.42249
Solar_radiation Soil_moisture | Precipitation NDVI	-0.33716
Temperature Soil_moisture | Precipitation NDVI Solar_radi...	0.013126

Current Editing Mode: Re-ordering Parent Nodes

Messages

Figure A.1: Anatomy of a Bayesian network (BN). This model is made in Uninet.

This appendix serves as a synopsis of the layout and terminology of a BN. The following items correspond with the numbers in Figure A.1.

1. **Variable:** a variable that has influence on the network, such as NDVI. Also a **node** in the BN. The bars in the node form a histogram of the values; the first number is the mean, and the number after the '±' is the standard deviation of the values. The unit of the variables is not mentioned in Uninet. For the inner workings of a BN, the unit does not matter.

2. **Target (variable):** the variable that is predicted in a test setup. Despite the fact that the probability distribution of all other variables is updated if one or more variables get conditioned, there is often one variable that is the final goal of a model; the variable for which the KGE is calculated. In this research the target variable in most of the tests is the monthly maximum daily average discharge (MMDAD).

3. **Connection:** also called a **(directed) edge**, an **arc**, an **arrow** or a **link** in this thesis. The number on the arrow denotes the **(conditional) normal rank correlation** (see Equation (2.12)). The direction of the arrow determines the inference between the variables in the model: the parent influences the child. This does not mean that a parent distribution cannot be updated from the child variable. This can be explained as follows: rain causes discharge, but when it is known that there is a high discharge, it can also be assumed that there was statistically more rain. The BN updates in both ways of the connection.

4. **Parent (variable):** for a variable, a parent is another variable on the tail end of the arrow which points to this variable. In this case the temperature is one of the parent variables of the soil moisture.

5. **Child (variable):** for a variable, a child is another variable on the head end of an arrow that starts at this variable. In this case the MMDAD is the child variable of the soil moisture. A parent variable is regarded as influencing the child variable in a BN.

6. **Order of dependencies:** this is the order of calculating the partial correlations in Equation (2.17). For the top variable, the correlation is just the normal rank correlation between the variables. For the variables lower in the order, this is the correlation of this variable and the child variable, given the correlation of all of the variables that are higher in this list and the child variable. In other words, it is the correlation of any information that is not already given by the variables higher in the order. This order changes the correlation of the connections and therefore the model workings. In this research, a mostly physically-based approach is taken to determine this order. The variables that are physically most directly related to the child variable are highest in the order. Other approaches would be a a solely performance-based order or a random order (such that model is easier to implement for the user).

7. **Fixing:** fixing one or more variables means setting this variable to a single known value. In Figure A.1 only the precipitation is fixed. This updates the distribution of the other variables via Equation (2.18). This is visible in the other nodes: the grey histograms are of the unconditioned histograms, the conditioned, updated distribution is shown in black.

B

Other methods to model a Bayesian network with copulas

B.1. Vine-copulas

Copulas are a model that represents the combined probability of multiple variables. For Archimedean copulas, all the variables, all variables are connected with the same parameter. Gaussian copulas, in contrast, can have multiple parameters. In a Bayesian network (BN), variables can have different correlations and dependencies. Therefore, a complex handling of copulas must be used to use Archimedean copulas in a BN. One of these methods is using vine-copulas (Appendix B.1) and another is using only one type of copula, which is split into using one Archimedean copula (Appendix B.2) or using one Gaussian copula with the conditional multivariate normal distribution (Section 2.6).

Vines were introduced by Bedford and Cooke (2001). The two most important ones are the Canonical (C-) vine and the D-vine. The C-vine is for variables that are connected in series and is defined as follows (Aas et al., 2009):

$$\begin{aligned} f(v_1,\dots,v_n) = \prod_{k=1}^{n} f(v_k) \prod_{j=1}^{n-1} \prod_{i=1}^{n-j} c_{i,(i+j)|(1,\dots,j-1)} \\ \left(F(v_j|(v_{i+1},\dots,v_{j-1})), F(v_{i+j}|(v_{i+1},\dots,v_{j-1}))\right). \end{aligned} \tag{B.1}$$

D-vines are constructed for variables connected in parallel from a single common variable and are defined as follows (Aas et al., 2009):

$$\begin{aligned} f(v_1,\dots,v_n) = \prod_{k=1}^{n} f(v_k) \prod_{j=1}^{n-1} \prod_{i=1}^{n-j} c_{i,(i+j)|(i+1,\dots,i+j-1)} \\ \left(F(v_i|(v_{i+1},\dots,v_{i+j-1})), F(v_{i+j}|(v_{i+1},\dots,v_{i+j-1}))\right). \end{aligned} \tag{B.2}$$

The number of vines per number of nodes n (variables) is

$$\frac{n!}{2}, \tag{B.3}$$

(Aas et al., 2009). For 8 variables, such as proposed in Table 4.1, this means that there are 20160 vines that have to be calculated, which is highly undesirable.

B.2. Multivariate Archimedean copulas

To avoid having to use copula-vines, it is also possible to use one multivariate copula for the whole BN. This is unfavourable because Archimedean copulas are not all defined easily in the multivariate form, and many have a tail dependence, which is likely to not perform well for all relations.

C

Derivations and additional equations

From Section 3.3.2:

The discharge Q is measured by multiplying the average flow velocity V_{avg} and the cross section times a factor (cA):

$$Q = cA \cdot V_{avg}. \tag{C.1}$$

The cross section multiplied by the factor is defined as follows from the four factors $c_0 \ldots c_3$:

$$cA = c_0 + c_1 h + c_2 h^2 + c_3 h^3, \tag{C.2}$$

where h is the water level. The average velocity V_{avg} is defined from n multiple velocity measurements that are evenly divided over the width of the stream:

$$V_{avg} = \frac{1}{n} \sum_{i=1}^{n} v_n. \tag{C.3}$$

FROM EQUATION (6.5):

$$
\begin{aligned}
f_{fit,un}(v) &= \frac{d}{dv} F_{fit,un}(v) \\
&= \frac{d}{de^{a_1(P(v)-a_0)}} \left(\frac{1}{1+e^{a_1(P(v)-a_0)}} \right) \\
&\cdot \frac{d}{da_1(P(v)-a_0)} \left(e^{a_1(P(v)-a_0)} \right) \\
&\cdot \frac{d}{dP(v)} (a_1(P(v)-a_0)) \\
&\cdot \frac{d}{dv} (P(v)) \;\; \text{[chain rule]} \\
&= \frac{-1}{\left(e^{a_1(P(v)-a_0)}+1\right)^2} \cdot e^{a_1(P(v)-a_0)} \cdot a_1 \cdot P'(v) \\
&= \frac{-a_1 \cdot P'(v) \cdot e^{a_1(P(v)-a_0)}}{\left(e^{a_1(P(v)-a_0)}+1\right)^2} \geq 0 \text{ for all } v \in (-\infty, +\infty)
\end{aligned}
\tag{C.4}
$$

FROM EQUATION (7.6):

$$
\begin{aligned}
KGE &= 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \\
\beta &= 1, KGE = 0.72 \\
0.72 &= 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (1-1)^2} = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2} \\
\sqrt{(r-1)^2 + (\alpha-1)^2} &= 0.28 \\
(r-1)^2 + (\alpha-1)^2 &= 0.28^2
\end{aligned}
\tag{C.5}
$$

D

Maps and figures

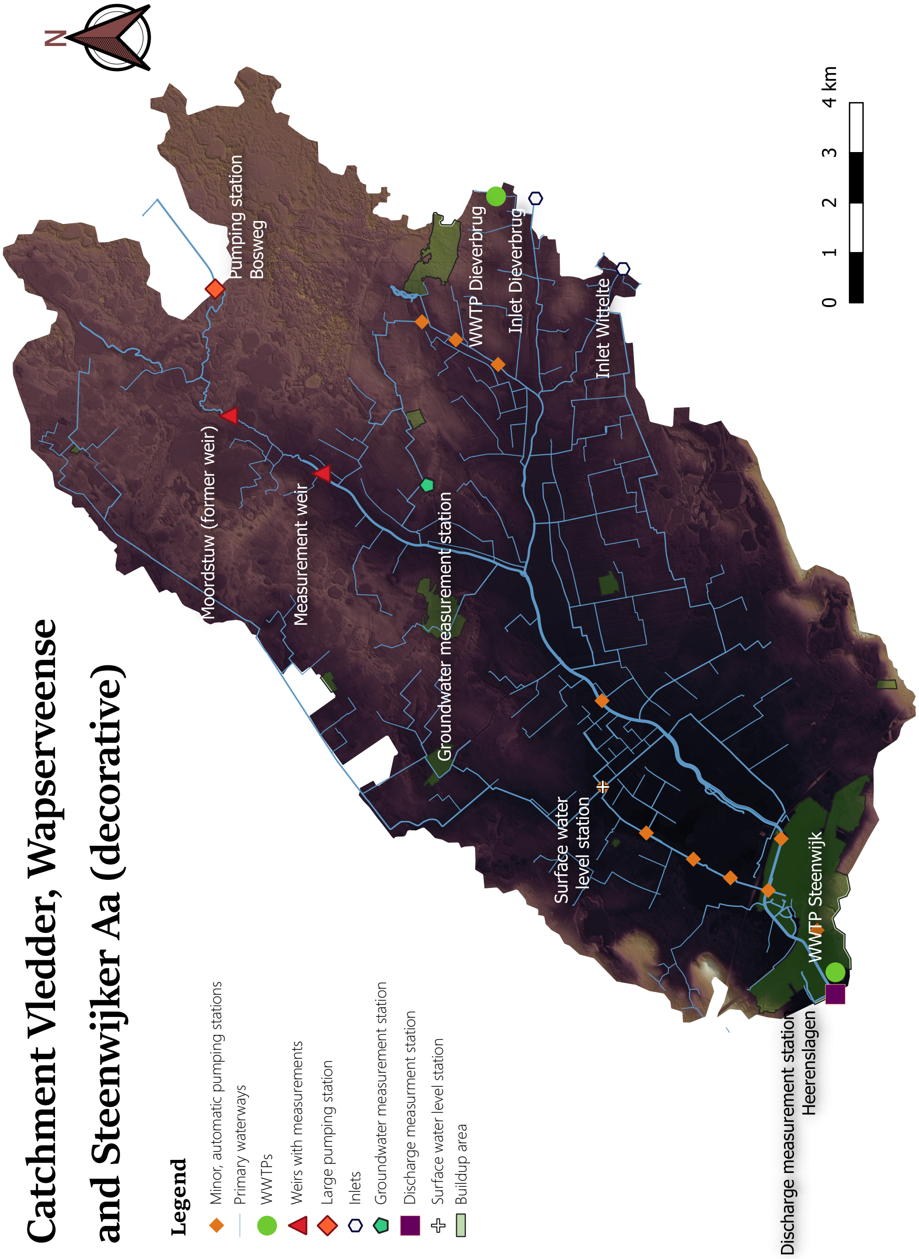

Catchment Vledder, Wapserveense
and Steenwijker Aa (decorative)
Legend
◆ Minor, automatic pumping stations
Primary waterways
● WWTPs
▲ Weirs with measurements
◆ Large pumping station
⬡ Inlets
⬢ Groundwater measurement station
■ Discharge measurment station
✛ Surface water level station
▯ Buildup area
N
Pumping station
Bosweg
Moordstuw (former weir)
Measurement weir
Groundwater measurement station
WWTP Dieverbrug
Inlet Dieverbrug
Inlet Wittelte
Surface water
level station
WWTP Steenwijk
Discharge measurement station
Heerenslagen
0 1 2 3 4 km

Figure D.1: Hillshaded catchment Vledder, Wapserveense and Steenwijker Aa

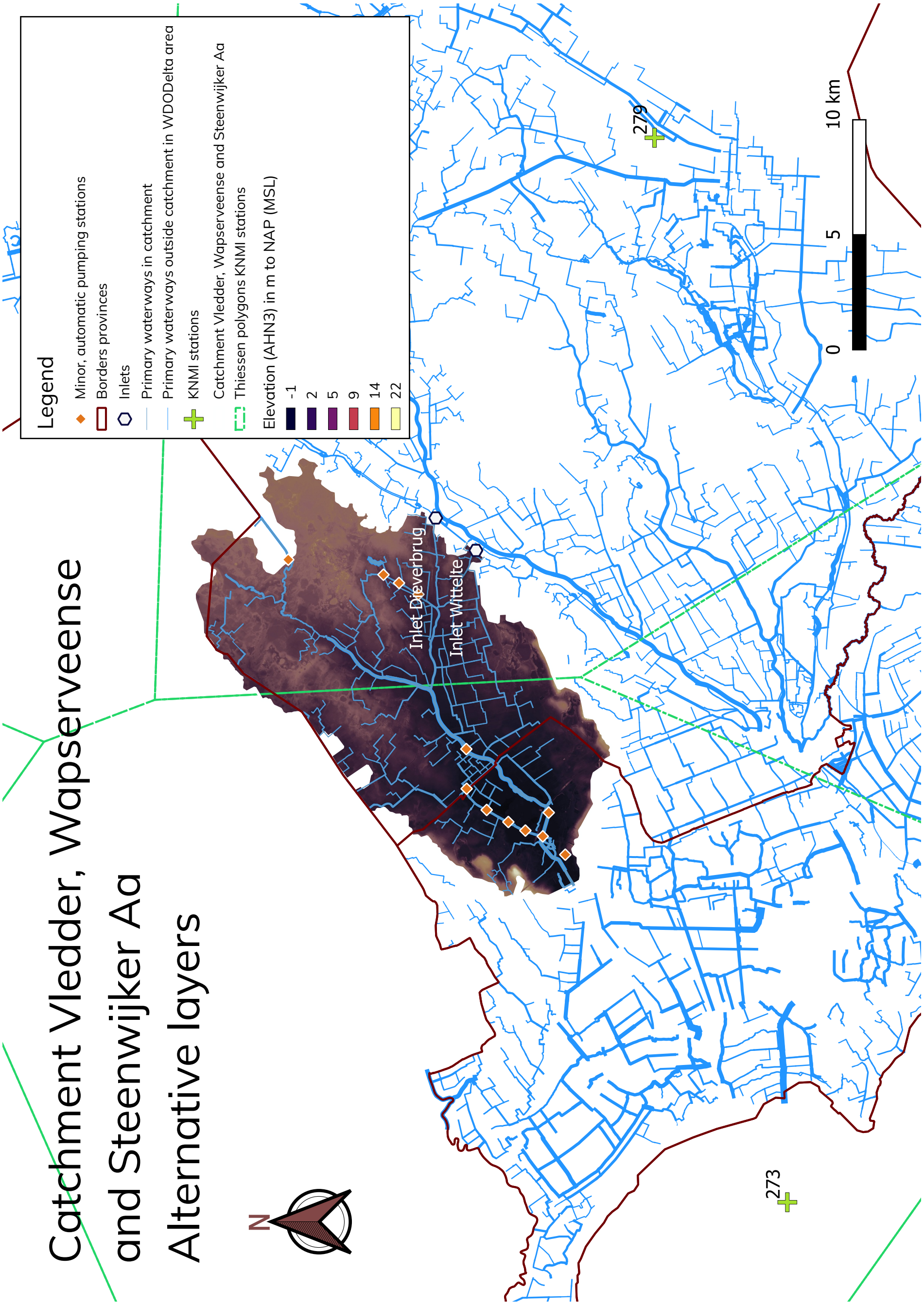

Catchment Vledder, Wapserveense
and Steenwijker Aa
Alternative layers
N
Legend
Minor, automatic pumping stations
Borders provinces
Inlets
Primary waterways in catchment
Primary waterways outside catchment in WDODelta area
KNMI stations
Catchment Vledder, Wapserveense and Steenwijker Aa
Thiessen polygons KNMI stations
Elevation (AHN3) in m to NAP (MSL)
-1
2
5
9
14
22
Inlet Dieverbrug
Inlet Wittelte
279
273
0
5
10 km

Figure D.2: Map of the catchment and its surroundings with different layers.

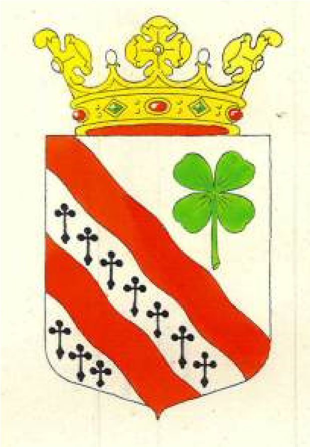


Figure D.3: Crest of the former Water Board "De Vledder en Wapserveense Aa", since 1959. The clover is a symbol for the newly planted agriculture. Source: Hoge Raad voor de Adel (1959)

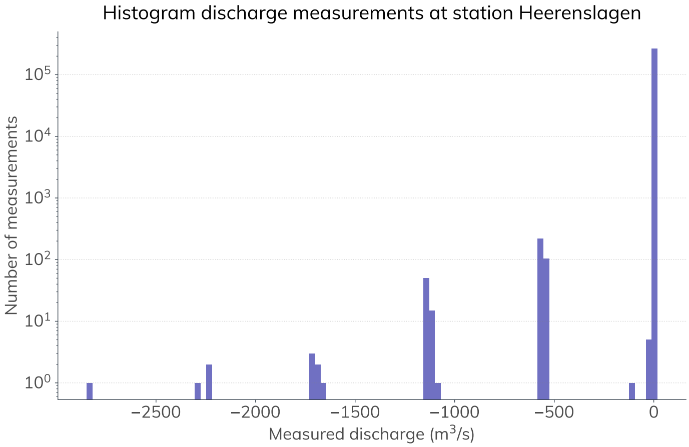


Figure D.4: Histogram of all the individual discharge measurements at Heerenslagen. Note the the y-axis is logarithmic.

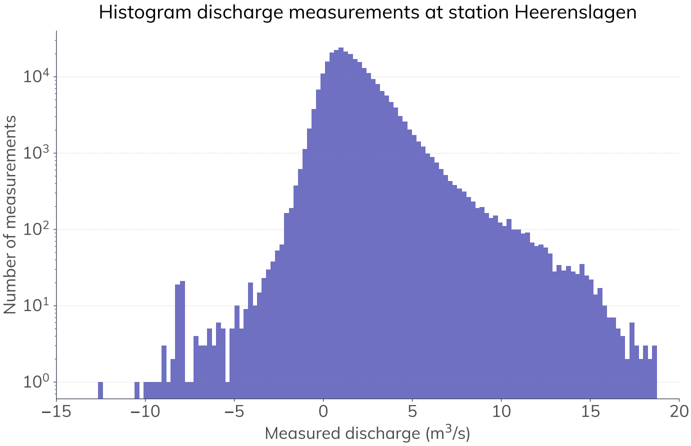


Figure D.5: Histogram of all the individual discharge measurements at Heerenslagen with a limited x-axis around the actual discharges. Note the the y-axis is logarithmic.

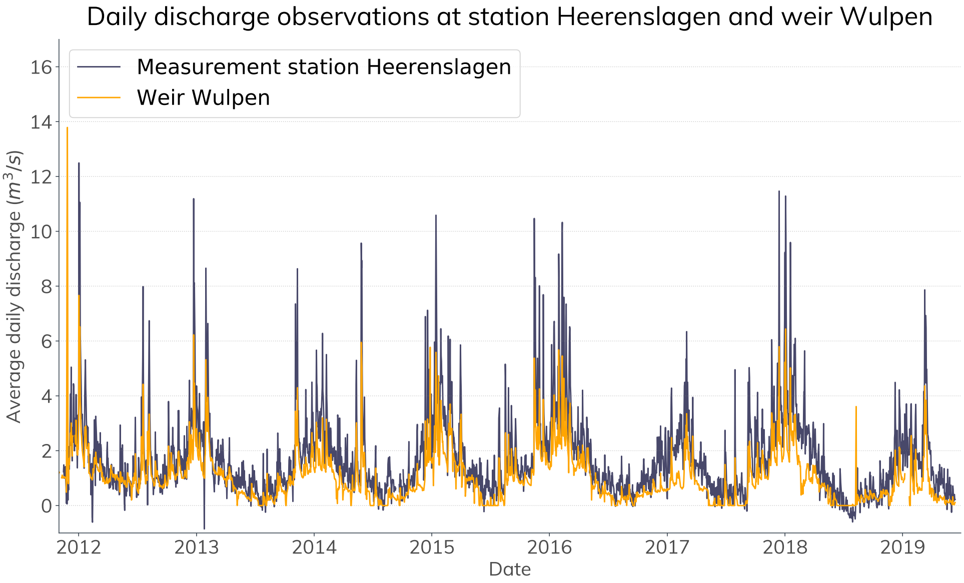
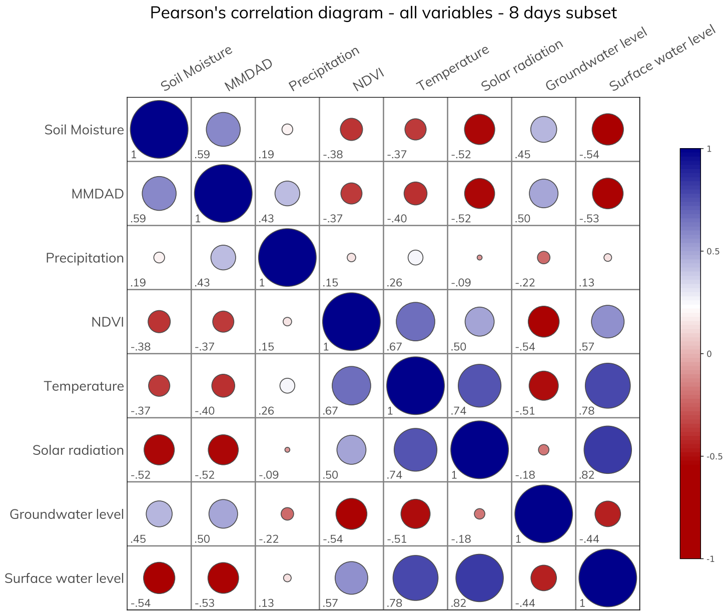
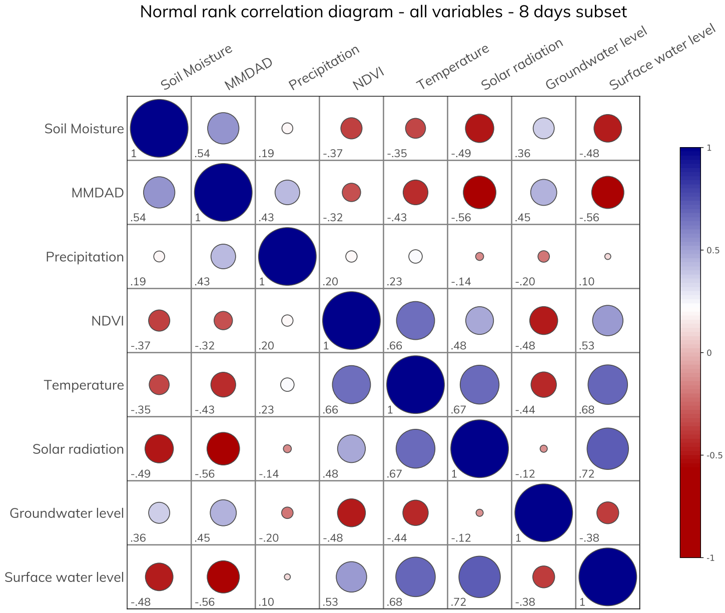


Figure D.6: Discharge measurements compared: measurement station Heerenslagen and weir Wulpen, which is 7 km upstream of the measurement station.



	Soil Moisture	MMDAD	Precipitation	NDVI	Temperature	Solar radiation	Groundwater level	Surface water level
Soil Moisture	1	.59	.19	-.38	-.37	-.52	.45	-.54
MMDAD	.59	1	.43	-.37	-.40	-.52	.50	-.53
Precipitation	.19	.43	1	.15	.26	-.09	-.22	.13
NDVI	-.38	-.37	.15	1	.67	.50	-.54	.57
Temperature	-.37	-.40	.26	.67	1	.74	-.51	.78
Solar radiation	-.52	-.52	-.09	.50	.74	1	-.18	.82
Groundwater level	.45	.50	-.22	-.54	-.51	-.18	1	-.44
Surface water level	-.54	-.53	.13	.57	.78	.82	-.44	1

Figure D.7: Correlation diagram of data used in this research. This diagram uses Pearson's correlation coefficient (see Section 2.4.2).



	Soil Moisture	MMDAD	Precipitation	NDVI	Temperature	Solar radiation	Groundwater level	Surface water level
Soil Moisture	1	.54	.19	-.37	-.35	-.49	.36	-.48
MMDAD	.54	1	.43	-.32	-.43	-.56	.45	-.56
Precipitation	.19	.43	1	.20	.23	-.14	-.20	.10
NDVI	-.37	-.32	.20	1	.66	.48	-.48	.53
Temperature	-.35	-.43	.23	.66	1	.67	-.44	.68
Solar radiation	-.49	-.56	-.14	.48	.67	1	-.12	.72
Groundwater level	.36	.45	-.20	-.48	-.44	-.12	1	-.38
Surface water level	-.48	-.56	.10	.53	.68	.72	-.38	1

Figure D.8: Correlation diagram of data used in this research. This diagram the normal rank correlation coefficient (see Section 2.4.2).

E

So..., is a Bayesian network machine learning?

According to Cambridge English Dictionary (2020), machine learning is defined as:

> The process of computers changing the way they carry out tasks by learning from new data, without a human being needing to give instructions in the form of a program

So, is a Bayesian network machine learning?

Yes, it is, because at least in the method proposed in this thesis, the computer changes the way it carries out tasks by learning from data, without human interference. In BNs, this happens with the construction of the correlation matrix and the empirical cumulative distribution function (ECDF), either fitted or not fitted. However, an argument could also be made for the opposite:

No, it is not, because BNs do not fit into the category of models that are generally (for example in scientific literature) regarded as machine learning models, such as naive Bayes, K nearest neighbours, K means, support vector machines and neural networks.

So, in my opinion, a sound argument could be made for Bayesian networks being machine learning or not. In general, I would advice a BN to be called a statistical method to scientist, because calling it machine learning raises wrong expectations, but do call it machine learning when you have to sell or promote the method, because this is not incorrect and it is highly popular and still gaining popularity (Google Trends, 2020).

F

ADDITIONAL MODEL AND TEST RESULTS

Variables		Type of copula				
u_1	u_2	Gaussian	Gumbel-H.	Clayton	Frank	Joe
Soil Moisture	Discharge	0.0197	**0.0163**	0.0385	0.0177	0.0184
Soil Moisture	Precipitation	0.0125	0.0170	0.0116	**0.0116**	0.0233
$-1\cdot$ Soil Moisture	NDVI	0.0186	0.0232	0.0227	**0.0153**	0.0309
$-1\cdot$ Soil Moisture	Temperature	0.0250	0.0252	0.0256	**0.0163**	0.0337
$-1\cdot$ Soil Moisture	Solar radiation	0.0263	0.0336	0.0255	**0.0225**	0.0451
Soil Moisture	Groundwater level	0.0263	0.0213	0.0402	0.0256	**0.0168**
$-1\cdot$ Soil Moisture	Surface water level	0.0297	0.0323	0.0327	**0.0205**	0.0432
Discharge	Precipitation	0.0126	0.0159	0.0201	**0.0125**	0.0241
$-1\cdot$ Discharge	NDVI	0.0146	0.0200	0.0157	**0.0129**	0.0276
$-1\cdot$ Discharge	Temperature	0.0156	0.0195	0.0264	**0.0122**	0.0269
$-1\cdot$ Discharge	Solar radiation	0.0147	0.0220	**0.0102**	0.0146	0.0342
Discharge	Groundwater level	**0.0143**	0.0151	0.0282	0.0144	0.0200
$-1\cdot$ Discharge	Surface water level	0.0144	0.0165	0.0314	**0.0118**	0.0228
Precipitation	Surface water level	**0.0135**	0.0143	0.0162	0.0139	0.0153
NDVI	Temperature	0.0267	0.0235	0.0477	**0.0170**	0.0320
NDVI	Solar radiation	0.0152	0.0200	0.0268	**0.0146**	0.0273
$-1\cdot$ NDVI	Groundwater level	0.0201	0.0237	0.0308	**0.0143**	0.0329
NDVI	Surface water level	0.0171	0.0234	0.0256	**0.0123**	0.0349
Temperature	Solar radiation	0.0228	0.0223	0.0470	**0.0168**	0.0294
$-1\cdot$ Temperature	Groundwater level	0.0204	0.0223	0.0319	**0.0153**	0.0293
Temperature	Surface water level	0.0230	0.0239	0.0453	**0.0147**	0.0335
$-1\cdot$ Solar radiation	Groundwater level	0.0217	0.0202	0.0303	**0.0196**	0.0198
Solar radiation	Surface water level	0.0269	0.0277	0.0476	**0.0169**	0.0377
$-1\cdot$ Groundwater level	Surface water level	0.0208	0.0262	0.0220	**0.0182**	0.0347
Average $\mathfrak{S}_{RMSD}$		0.01969	0.02190	0.02918	**0.01589**	0.0289

Table F.1: $\mathfrak{S}_{RMSD}$ values (–) of the altered Cramér-von Mises tests for all of the connections in the Bayesian network of the first model iteration. The lowest value for each combination of variables is shown in **bold**.

Variables		Type of copula			
u_1	u_2	Gaussian	Gumbel-H.	Clayton	Frank
Soil Moisture	Discharge	0.018	**0.140**	0.000	0.046
Soil Moisture	Precipitation	**0.730**	**0.188**	**0.864**	**0.800**
$-1\cdot$ Soil Moisture	NDVI	**0.054**	0.014	0.016	**0.198**
$-1\cdot$ Soil Moisture	Temperature	0.004	0.006	0.000	**0.110**
$-1\cdot$ Soil Moisture	Solar radiation	0.002	0.000	0.004	0.004
Soil Moisture	Groundwater level	0.000	0.018	0.000	0.000
$-1\cdot$ Soil Moisture	Surface water level	0.000	0.000	0.000	0.022
Discharge	Precipitation	**0.622**	**0.238**	0.044	**0.640**
$-1\cdot$ Discharge	NDVI	**0.348**	0.044	**0.276**	**0.570**
$-1\cdot$ Discharge	Temperature	**0.224**	**0.056**	0.000	**0.690**
$-1\cdot$ Discharge	Solar radiation	**0.296**	0.012	**0.944**	**0.290**
Discharge	Groundwater level	**0.322**	**0.314**	0.000	**0.298**
$-1\cdot$ Discharge	Surface water level	**0.326**	**0.144**	0.000	**0.690**
Precipitation	Surface water level	**0.544**	**0.484**	**0.296**	**0.424**
NDVI	Temperature	0.000	0.002	0.000	**0.068**
NDVI	Solar radiation	**0.252**	0.036	0.004	**0.308**
$-1\cdot$ NDVI	Groundwater level	0.022	0.002	0.000	**0.266**
NDVI	Surface water level	**0.102**	0.006	0.000	**0.600**
Temperature	Solar radiation	0.000	0.004	0.000	**0.080**
$-1\cdot$ Temperature	Groundwater level	0.018	0.012	0.000	**0.168**
Temperature	Surface water level	0.004	0.002	0.000	**0.174**
$-1\cdot$ Solar radiation	Groundwater level	0.010	0.050	0.000	0.026
$-1\cdot$ Solar radiation	Surface water level	0.000	0.000	0.000	0.048
$-1\cdot$ Groundwater level	Surface water level	0.018	0.000	0.024	**0.066**
Average p-value		0.163	0.074	0.103	0.274

Table F.2: Average p-value for the $\mathfrak{S}_{RMSD}$ values of different bivariate copulas for all combinations in model 1, with the method of Genest and Remillard (2008), see Section 5.3, with $N = 500$. All copulas for combinations of parameters that are not rejected with $\alpha = 0.05$ are denoted in **bold**.

Variables		**Quadrant**				
u_1	u_2	NE	NW	SE	SW	All data
Soil Moisture	Discharge	0.273	0.144	**0.464**	0.089	0.369
Soil Moisture	Precipitation	**-0.139**	**0.177**	-0.029	0.007	0.036
Soil Moisture	NDVI	-0.440	-0.059	-0.148	0.044	-0.476
Soil Moisture	Temperature	-0.260	0.172	-0.040	**0.500**	-0.343
Soil Moisture	Solar radiation	-0.271	0.118	-0.377	0.194	-0.501
Soil Moisture	Groundwater level	0.182	-0.082	0.152	**-0.416**	0.290
Soil Moisture	Surface water level	-0.266	0.192	-0.023	**0.619**	-0.407
Discharge	Precipitation	0.096	-0.304	0.101	0.061	0.384
Discharge	NDVI	**-0.423**	-0.021	-0.214	-0.153	-0.286
Discharge	Temperature	0.134	-0.041	-0.019	0.057	-0.399
Discharge	Solar radiation	0.093	-0.011	-0.191	-0.210	-0.528
Discharge	Groundwater level	-0.147	**0.204**	**0.219**	**-0.417**	0.177
Discharge	Surface water level	0.168	-0.364	0.075	-0.295	-0.514
Precipitation	Surface water level	**0.307**	**0.358**	0.022	-0.004	0.244
NDVI	Temperature	0.393	0.003	**0.700**	0.135	0.641
NDVI	Solar radiation	-0.022	0.156	**0.615**	0.005	0.577
NDVI	Groundwater level	-0.031	-0.034	-0.208	-0.400	-0.441
NDVI	Surface water level	0.261	-0.155	0.025	0.030	0.616
Temperature	Solar radiation	0.404	0.221	-0.174	0.218	0.753
Temperature	Groundwater level	0.179	0.192	-0.067	0.025	-0.279
Temperature	Surface water level	0.169	-0.004	0.036	0.073	0.675
Solar radiation	Groundwater level	**0.316**	**0.128**	0.051	0.106	-0.123
Solar radiation	Solar radiation	0.056	-0.404	0.175	0.081	0.646
Groundwater level	Surface water level	-0.246	0.024	-0.212	-0.280	-0.408

Table F.3: Pearson's correlation values for the quadrants of the normalised data, for all combinations in model 1. Quadrants with a higher absolute correlation than the total correlation point in the direction of tail dependence and are shown in **bold**. The parameters are not multiplied by -1 in case of negative correlation, as this does not matter for this test. See also Section 5.5.1.

Variables		Type of copula																			
		Gaussian					Gumbel-H.					Clayton					Frank				
u_1	u_2	NE	NW	SE	SW	Avg.	NE	NW	SE	SW	Avg.	NE	NW	SE	SW	Avg.	NE	NW	SE	SW	Avg.
Soil Moisture	Discharge	-0.05	-0.02	0.32	-0.25	0.16	-0.28	0.09	0.39	-0.14	0.22	0.21	0.11	0.39	-0.37	0.27	0.05	0.08	0.41	-0.17	0.18
Soil Moisture	Precipitation	-0.24	0.15	-0.08	-0.08	0.14	-0.32	0.15	-0.06	-0.01	0.13	-0.21	0.12	-0.10	-0.30	0.18	-0.21	0.15	-0.05	-0.05	0.11
−1· Soil Moisture	NDVI	-0.15	0.21	0.19	-0.18	0.18	-0.34	0.32	0.24	-0.12	0.26	-0.02	0.28	0.23	-0.44	0.24	-0.07	0.25	0.21	-0.11	0.16
−1· Soil Moisture	Temperature	-0.42	0.10	-0.33	-0.14	0.25	-0.65	0.16	-0.26	-0.10	0.29	-0.30	0.11	-0.28	-0.40	0.27	-0.42	0.13	-0.28	-0.10	0.23
−1· Soil Moisture	Solar radiation	-0.55	0.19	-0.65	-0.08	0.37	-0.71	0.29	-0.54	0.02	0.39	-0.29	0.27	-0.57	-0.34	0.37	-0.46	0.26	-0.57	-0.02	0.33
Soil Moisture	Groundwater lvl.	-0.04	-0.20	0.03	-0.62	0.22	-0.27	-0.15	0.11	-0.60	0.28	0.11	-0.20	0.10	-0.58	0.25	0.05	-0.15	0.07	-0.57	0.21
−1· Soil Moisture	Surface water lvl.	-0.54	-0.25	-0.73	-0.10	0.41	-0.75	-0.15	-0.64	0.00	0.39	-0.34	-0.15	-0.66	-0.34	0.37	-0.50	-0.16	-0.67	-0.07	0.35
Discharge	Precipitation	-0.15	-0.46	-0.03	-0.17	0.20	-0.31	-0.34	0.07	-0.09	0.20	0.03	-0.40	0.03	-0.42	0.22	-0.06	-0.37	0.03	-0.13	0.15
−1· Discharge	NDVI	-0.12	0.35	0.05	0.10	0.16	-0.30	0.41	0.10	0.16	0.24	-0.04	0.38	0.07	-0.14	0.16	-0.07	0.38	0.08	0.15	0.17
−1· Discharge	Temperature	-0.10	0.05	0.00	-0.23	0.09	-0.31	0.09	0.12	-0.14	0.16	0.05	0.09	0.05	-0.38	0.14	-0.05	0.11	0.05	-0.16	0.09
−1· Discharge	Solar radiation	-0.38	-0.54	-0.29	0.17	0.34	-0.57	-0.43	-0.21	0.33	0.38	-0.10	-0.44	-0.22	-0.11	0.22	-0.28	-0.47	-0.22	0.29	0.31
Discharge	Groundwater lvl.	-0.40	0.06	0.06	-0.71	0.31	-0.60	0.13	0.18	-0.56	0.37	-0.22	0.13	0.11	-0.84	0.32	-0.30	0.10	0.17	-0.59	0.29
−1· Discharge	Surface water lvl.	-0.24	-0.30	0.11	-0.56	0.31	-0.45	-0.23	0.20	-0.44	0.33	-0.03	-0.26	0.21	-0.68	0.30	-0.17	-0.21	0.16	-0.46	0.25
Precipitation	Surface water lvl.	0.24	0.31	0.00	-0.05	0.15	0.16	0.32	-0.01	-0.03	0.13	0.25	0.14	0.22		0.20	0.27	0.32	-0.01	-0.02	0.15
NDVI	Temperature	-0.04	-0.19	0.52	-0.32	0.27	-0.29	-0.06	0.64	-0.20	0.30	0.32	-0.08	0.61	-0.48	0.37	-0.01	-0.10	0.64	-0.25	0.25
NDVI	Solar radiation	-0.32	0.02	0.46	-0.29	0.27	-0.45	0.12	0.58	-0.15	0.33	-0.07	0.08	0.56	-0.44	0.29	-0.19	0.08	0.55	-0.18	0.25
−1· NDVI	Groundwater lvl.	-0.30	0.08	0.06	-0.18	0.15	-0.54	0.17	0.13	-0.04	0.22	-0.09	0.12	0.13	-0.37	0.18	-0.23	0.18	0.15	-0.09	0.16
NDVI	Surface water lvl.	-0.08	-0.32	-0.12	-0.29	0.20	-0.25	-0.19	-0.04	-0.18	0.16	0.22	-0.23	-0.02	-0.49	0.24	0.02	-0.24	-0.07	-0.19	0.13
Temperature	Solar radiation	-0.07	0.03	-0.38	-0.26	0.18	-0.26	0.15	-0.20	-0.14	0.18	0.35	0.18	-0.24	-0.35	0.28	0.03	0.18	-0.26	-0.17	0.16
−1· Temperature	Groundwater lvl.	-0.21	0.08	0.10	-0.21	0.15	-0.39	0.12	0.16	-0.10	0.19	0.00	0.11	0.16	-0.38	0.16	-0.14	0.11	0.16	-0.12	0.13
Temperature	Surface water lvl.	-0.31	-0.15	-0.17	-0.42	0.26	-0.49	-0.06	-0.08	-0.29	0.23	0.13	-0.10	-0.03	-0.56	0.20	-0.22	-0.05	-0.07	-0.34	0.17
−1· Solar radiation	Groundwater lvl.	-0.23	-0.21	-0.19	-0.36	0.25	-0.39	-0.17	-0.17	-0.36	0.27	-0.16				0.16	-0.21	-0.17	-0.18	-0.32	0.22
−1· Solar radiation	Solar radiation	-0.43	-0.57	0.00	-0.42	0.36	-0.63	-0.48	0.11	-0.30	0.38	-0.01	-0.46	0.08	-0.55	0.28	-0.37	-0.48	0.13	-0.36	0.34
−1· Groundwater lvl.	Surface water lvl.	-0.23	0.19	0.13	0.04	0.15	-0.41	0.29	0.22	0.11	0.26	-0.12	0.26	0.18	-0.23	0.20	-0.17	0.24	0.22	0.12	0.19
Average						0.23					0.26					0.25					0.21

Table F.4: Differences in Pearsons's correlation coefficient per quadrant between data points and 20000 samples generated from fitted copula. A negative number denotes that the actual correlation was higher for this quadrant, than the correlation from the generated samples. Averages are taken from absolute values. See Section 5.5.2. The empty spots lacked points in these quadrants

G

Additional information

G.1. Language report

British English

G.2. Hardware used

PC	HP Probook 650 G2
Processor	Intel^R Core^{TM} i5-6200U CPU @2.30 GHz 2.40 GHz
RAM	16.0 GB (15.9 GB usable)

G.3. Software used

G.3.1. General software

Operating system	Windows 10 Professional
Programming	Python 3.6 and Python 3.7 (Anaconda release)
Python editor	Anaconda Spyder 3.3.5 and Spyder 4.0.1
Bayesian network	Uninet[1]
GIS	QGIS 3.81 with GRASS 7.6.1
LaTeXeditor	Overleaf
LaTeXcompiler	XeLaTeX
SOBEK	SOBEK 213
Additional text editor	Atom 1.46

G.3.2. Python modules used

- `copulabayesnet` (see Appendix G.4)
- `numpy`
- `pandas`

[1] Available on `https://lighttwist-software.com/uninet/`

- `matplotlib`
- `pycopula`
- `scipy`
- `datetime`
- `rasterio`
- `statsmodels`
- `sklearn.metrics`
- `geopandas`
- `pingouin`
- `easygui`
- `biokit`
- `sys`
- `os`
- `mpl_toolkits`

G.4. PROGRAMMES WRITTEN FOR THIS THESIS

For this thesis, a Python package called `copulabayesnet`, has been developed.
It is distributed with the MIT licence and can be accessed through `https://github.com/SjoerdGn/copulabayesnet` or `https://pypi.org/project/copulabayesnet/` and can be directly installed by entering `pip install copulabayesnet` in the command prompt. The Python package contains all methods of the copula testing and the full multivariate normal method, including methods to test it k-fold over a whole dataset, as well as methods to plot copulas and results. For access to the data, please contact Waterschap Drents Overijsselse Delta and the author, for the supporting code, please contact Witteveen+Bos and the author.