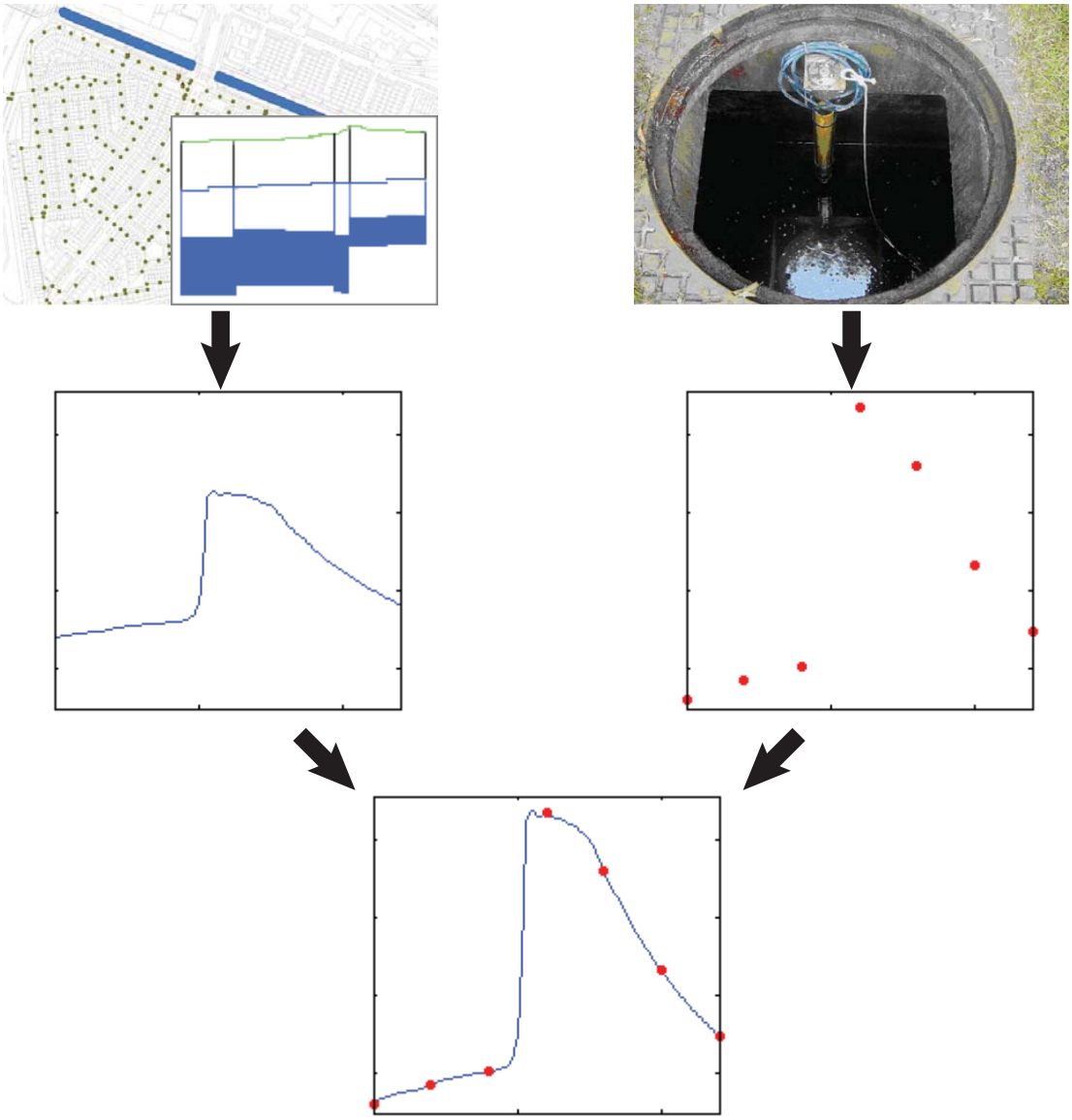
Combining Field Observations and Hydrodynamic Models in Urban Drainage

Design of a Monitoring Network and the Application of Data Assimilation



Johan Post

October 2012

TU Delft Delft University of Technology

Challenge the future

Combining field observations and hydrodynamic models in urban drainage

Design of a monitoring network and the application of data assimilation

Johan Post

for the degree of:

Master of Science in Civil Engineering

Date of submission: 18 October 2012
Date of defence: 25 October 2012

Committee:

Prof.dr.ir. F.H.L.R. Clemens	Delft University of Technology Sanitary Engineering Section
Prof.dr.ir. A.W. Heemink	Delft University of Technology Mathematical Physics
Dr.ir. J.L. Korving	Delft University of Technology Mathematical Physics

Sanitary Engineering Section, Department of Water Management
Faculty of Civil Engineering and Geosciences
Delft University of Technology, Delft

Abstract

Hydrodynamic models are often the main source of information used in the engineering practice to judge the performance of urban drainage systems with respect to the occurrence of flooding and the occurrence of spills of wastewater into open water courses. If the hydraulic performance is deemed insufficient, these models are used to determine the effect of proposed alterations to the system.

Due to the fact that hydrodynamic models are subject to uncertainties originating from various sources, the reliability of the model results is limited. These uncertainties can be quantified when a calibrated model is applied. However, is has been found that water level predictions for a storm event other than the storm event used to calibrate the model are less reliable. In other words, the portability of parameter sets obtained in calibration based on single storm events is limited. Logically, the use of continuous time series containing multiple storm events for model calibration does not result in an equally good match for a specific storm event, compared to single event calibration.

This research is aimed at investigating whether data assimilation can be applied to models in urban drainage in order to simulate field observations for continuous time series. The process of data assimilation combines measurements and models by updating the set of model parameter values when new measurements are available (see Figure 1) . In order to obtain sufficient field observations for data assimilation, a method for the design of a monitoring network capable of collecting information on the relevant processes is elaborated.

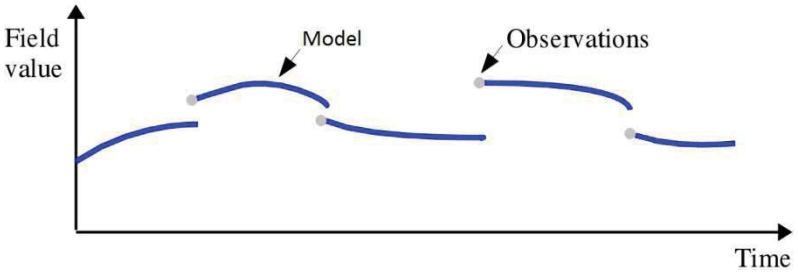


Figure 1: principle of data assimilation, adapted from (Solonen, 2011)

For two simple examples, the applied data assimilation method is able to accurately simulate water levels and proves to be robust with respect to the initial estimation of the parameter values. Due to time constraints, the anticipated implementation of the data assimilation method for the case study of Delft's city centre drainage system has been unsuccessful. For the same case study, a monitoring network has been designed. This network is able to collect a wide variety of information with respect to the parameters of interest, while still incorporating some form of redundancy to account for sensor failure and for the cross validation of data.

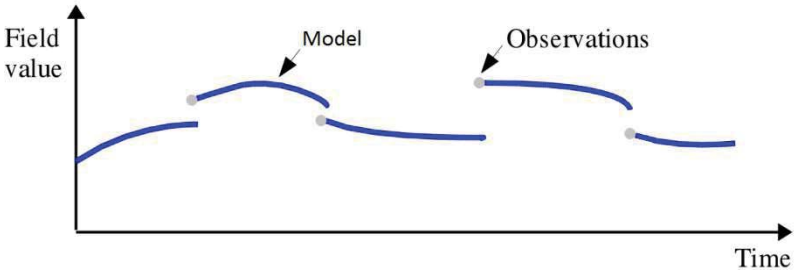
Although it has been found that data assimilation is successful in simulating water levels for continuous time series, more research is needed on the implementation for large scale application in the field of urban drainage. Due to the spread of information over the potential monitoring locations, a well-designed monitoring network is considered to be a prerequisite for the collection of sufficient information for the application of data assimilation.

Samenvatting

Hydrodynamische modellen zijn een belangrijke bron van informatie om het hydraulisch functioneren van rioleringssystemen te beoordelen, en daarmee inzicht te verkrijgen in de gevoeligheid voor wateroverlast en het optreden van riool overstorten. Als het functioneren onvoldoende wordt beoordeelt, kan het effect van aanpassingen aan het rioleringssysteem berekend worden.

Door onzekerheden vanuit verschillende bronnen, is de betrouwbaarheid van deze modellen beperkt. Deze onzekerheden kunnen gekwantificeerd worden door een gekalibreerd model te gebruiken. Het is echter gebleken, dat voorspellingen voor een andere neerslaggebeurtenis dan gebruikt is om het model te kalibreren minder betrouwbaar zijn. In andere woorden, de overdraagbaarheid van de set model parameters verkregen door middel van kalibratie voor een neerslaggebeurtenis is beperkt. Het gebruik van een neerslagreeks voor het kalibreren van een model resulteert niet in een even goede match in vergelijking met de kalibratie voor één neerslaggebeurtenis.

Deze scriptie heeft als doel om te onderzoeken of data assimilatie toegepast kan worden op rioleringsmodellen om veldwaarnemingen te simuleren voor tijdreeksen. Data assimilatie omvat het combineren van metingen en modellen door de set van model parameters te actualiseren als nieuwe metingen beschikbaar komen (zie ook Figuur 2). Om voldoende veldwaarnemingen te verkrijgen voor data assimilatie, wordt een methode uitgewerkt om een meet netwerk te ontwerpen dat in staat is voldoende informatie te verzamelen over de relevante parameters.



Figuur 2: principe van data assimilatie, aangepast van (Solonen, 2011)

Voor twee simpele voorbeelden, is de toegepaste data assimilatie methode in staat om nauwkeurig waterstanden te simuleren en is de methode robuust met betrekking tot de eerste schatting van de parameter waarden. Vanwege de beperkte tijd is de verwachte implementatie van de data assimilatie methode voor de case studie niet succesvol. Voor dezelfde case studie is een meetnetwerk ontworpen. Dit meetnetwerk is in staat om informatie te verzamelen met betrekking tot de relevante parameters, terwijl er voldoende overlap in de verzamelde data is om rekening te houden met sensor defecten en voor de cross validatie van data.

Hoewel het is gebleken dat data assimilatie succesvol waterstanden kan simuleren voor tijdreeksen, is er meer onderzoek nodig naar de implementatie van data assimilatie voor grootschalige toepassing in de riolering. Als gevolg van de spreiding van informatie over de potentiele meetlocaties, wordt een goed ontworpen meet netwerk beschouwd als een voorwaarde voor het verzamelen van voldoende informatie om data assimilatie toe te passen.

Preface

The thesis lying before you is the result of 9 months of research at the Technical University of Delft. My interest in hydrodynamic modelling originates from earlier work at Waternet, and knowledge obtained at the University. I hope my work has made a small contribution to the field of urban drainage, and more important inspires further research.

As is common, I would like to take the opportunity to thank several people who all in their own way helped me with my graduation work. In a somewhat chronological sequence I would like to start by thanking my family for their unconditional support. Life throughout my Master would not have been the same without the humor and help of my fellow students Bram Stegeman and Paul Buring. Furthermore my thanks goes out to the Rekengroep.

I would like to thank my graduation committee consisting of François Clemens, Hans Korving and Arnold Heemink for their wisdom, expertise and patience. François, I consider the weekly discussions we had to be one of the most constructive parts of my education in Delft, and your enthusiasm really inspired me throughout these past months. I'm looking forward to continue this collaboration for my PhD.

My appreciation goes out to Deltares for lending me a Sobek license for the duration of my thesis. I would also like to extend my gratitude to Nils van Velzen, Martin Verlaan and Edwin Bos for their help with OpenDA. Edwin, without you're tips and guidance I would not have reached the Java skill level needed, in the time available.

Content

1 Introduction 15
1.1 Problem 15
1.2 Research aim 17
1.3 Possibilities for data assimilation 18
1.4 Outline of the thesis 19

2 Literature review 21
2.1 Significance of model calibration 21
2.2 Deficiencies of static calibration 21
2.3 Dynamic calibration for time series 21
2.4 Measurements for model calibration 22

3 Case study characteristics 25

4 Monitoring network 29
4.1 Information requirement 29
4.2 Design of a monitoring network 30
4.2.1 Methods for water quantity measurements 30
4.2.2 Measuring frequency and accuracy 31
4.2.3 Spatial distribution of the monitoring locations 33
4.2.3.1 Excluding locations 34
4.2.3.2 Sensitivity analysis 34
4.2.3.3 Singular value decomposition 35
4.2.3.4 Optimisation of the information content 36
4.2.4 Sequence of operations 40

5 Data assimilation 43
5.1 The Kalman filter 43
5.2 The ensemble Kalman filter 45
5.2.1 Example of the EnKF for a simple reservoir model 47

6 Measuring setup case study 51
6.1 Omitted locations 51
6.2 Genetic algorithm for de-correlation 51
6.3 Parameters for optimization 59
6.4 Monitoring locations 63
6.4.1 Sensor correlation 64
6.5 Measuring frequency 65

7 Data assimilation in urban drainage modelling 69
7.1 Two node model with constant inflow 70
7.2 Delft city centre case study 71
7.2.1 Analysis of the results 73

8 Conclusions and recommendations 75
8.1 Conclusions 75
8.2 Recommendations 76

List of references 79

Annexe I: Singular values and Eigenvectors for the weirs 85

Annexe II: Applied storm events **87**

Annexe III: Singular values and Eigenvectors for the applied storms **91**

Annexe IV: Layout of the monitoring network **93**

Annexe V: Kalman filter example **95**

Annexe VI: OpenDA structure **99**

Annexe VII: XML configuration **101**

List of illustrations

Figure 1: principle of data assimilation, adapted from (Solonen, 2011) 3
Figuur 2: principe van data assimilatie, aangepast van (Solonen, 2011) 5
Figure 3: Pluvial flooding during a storm event in Amsterdam, 2007 15
Figure 4: Integrated Sewer System Management Process (Nederlands Normalisatie Instituut, 2008) 15
Figure 5 : modelling a second degree polynomial with first degree polynomials 16
Figure 6: Principle of data assimilation 17
Figure 7: Management operations adapted from (Nederlands Normalisatie Instituut, 1994) 18
Figure 8: Thesis structure 19
Figure 9: Uncertainty bandwidth in model predictions (20^{th} and 80^{th} percentile) and validation data after calibration on the left and after 2^{nd} update on the right (Rauch, et al., 2011)22
Figure 10: Total information content during two rainfall events (Henckens & Clemens, 2004)23
Figure 11: Aerial photo of Delft's city centre (Municipality of Delft, 2012) 25
Figure 12: schematic representation of the relevant sewer districts 25
Figure 13: Canal in the city centre of Delft (Dijk, van, Z., 2005) 26
Figure 14: Overview of the engineering works used to create the isolated water system 27
Figure 15: Adjustable weir in closed position 27
Figure 16: Scenario where the sewer system transports water from one surface water body to another 27
Figure 17:The knowledge household (Lohuizen, van, C.W.W., 1986) 29
Figure 18: Monitoring cycle (UN/ECE Task Force on Monitoring & Assessment, 2000) 30
Figure 19: example of a 50 and 120 Hz sinusoid in the time domain and frequency domain (Mathworks, 2012) 31
Figure 20: Example of a power spectral density function (PSD) 32
Figure 21: difficult accessible manhole 34
Figure 22: Schematic representation of singular values and eigenvectors for a ellipse (Lay, 2006) 36
Figure 23: weight factor for different overlap values 38
Figure 24: weight factor for different values of the standard deviation and $\mu = 0.8$ 39
Figure 25: Correlation between multiple sensors and correlation between two sensors 40
Figure 26: Working sequence for the design of a monitoring network 41
Figure 27: The application of sequential data assimilation on the state with respect to observations (Eskes, et al., 1998) 43
Figure 28: Kalman filter steps; the time step is denoted by k and the forecast by f. 44
Figure 29: EnKF steps for an ensemble size i = 3 46
Figure 30: Reservoir model 47
Figure 31: Principle of the twin model data assimilation concept 48
Figure 32: application of the EnKF on the reservoir model 49
Figure 33: Types of area identified to be less suitable for the placement of monitoring equipment (also see the above description) 51
Figure 34: empirical cumulative distribution function for 100 locations 52
Figure 35: increasing antibiotic resistance due to natural selection (Urbano , 2010) 53
Figure 36: Shape of a objective function for a single parameter (Sparknotes, 2012) 53
Figure 37: Genetic algorithm flowchart 54
Figure 38: creation of the next generation by three kinds of children (Mathworks, 2007) 54
Figure 39: example of the total information content for two monitoring locations 55
Figure 40: mean of the total information content using the standard mixed integer optimizer56

Figure 41: standard deviation of the total information content using the standard mixed integer optimizer 56
Figure 42: mean of the total information content using the custom optimizer 57
Figure 43: standard deviation of the total information content using the custom optimizer 58
Figure 44: Mean information content of 31 monitoring locations chosen from 1045 locations where sensors can be installed 59
Figure 45: Standard deviation of 31 monitoring locations chosen from 1045 locations where sensors can be installed 59
Figure 46: Typical inverted siphon construction under a city canal with a CSO on both sides 60
Figure 47: Values of the eigen vector for different singular values 62
Figure 48: Spatial distribution of the 12 sensors derived from the optimisation algorithm 64
Figure 49: Box and Whisker plot of the correlations coefficients for the derived monitoring locations 65
Figure 50: Singular values normalized with respect to Δt = 1 min for 27 different parameters 66
Figure 51: Cumulative distribution function of the relative information content, with enlargement of the top part 67
Figure 52: Twin model concept for Sobek and OpenDA 69
Figure 53: Difference between a normal simulation and a EnKF simulation for an arbitrary model 70
Figure 54: schematisation of the Sobek model containing one link and two computational nodes 70
Figure 55: Observations and model predictions of the EnKF 71
Figure 56: Setup used to transfer the state after each analysis step 72
Figure 57: Divergence of the optimisation method due to 72
Figure 58: Difference in calculated water depth when the model is run in several parts 73
Figure 59: Cumulative precipitation for when the storm event is simulated in one run and in six sub runs 73
Figure 60: simulated water depth for one of the monitoring locations compared to the observations 74
Figure 61: abbreviated schematisation of the optimisation process 75
Figure 62: Standard design storm 02 with a return period of 0.25 years 87
Figure 63: Historical storm 19-01-12 88
Figure 64: Historical storm 24-07-11 89
Figure 65: diagram of the xml build of OpenDA 99
Figure 66: an example of Gaussian white noise added to observations produced by a model 100

List of tables

Table 1: Characteristics of the Delft city centre sewer system 26
Table 2: Set of parameters relevant to the applied model after clustering the weir coefficients based on identifiability 61
Table 3: Singular values and corresponding Eigen vectors for 15 locations selected by the genetic algorithm 62
Table 4: Selection of the singular values and corresponding eigen vectors determined by the DWF and channel friction 63
Table 5: Eigen vectors dominated by weir coefficients in descending order with respect to the singular values, Δt = 3 min 67
Table 6: Eigen vectors dominated by weir coefficients in descending order with respect to the singular values, Δt = 4 min 68
Table 7: Storm events with a precipitation sum between the 15 and 25 mm 88

List of abbreviations

CSO	**C**ombined **S**ewer **O**verflow
DA	**D**ata **A**ssimilation
DWF	**D**ry **W**eather **F**low
EKF	**E**xtended **K**alman **F**ilter
EnKF	**En**semble **K**alman **F**ilter
KF	**K**alman **F**ilter
NAP	**N**ormaal **A**msterdams **P**eil, or Amsterdam Ordnance Datum

1 Introduction

One of the main objectives of sewer systems is to collect and transport excess storm water. When the capacity of a drainage system is insufficient during a storm event, not all the water is transported from the surface. This is referred to as pluvial flooding (see Figure 3).

Figure 3: Pluvial flooding during a storm event in Amsterdam, 2007

The design of a drainage system is an optimisation between economic benefits like the prevention of damage and nuisance, and the costs involved in constructing and maintaining the system. From this process of optimisation a standard, often related to a return period is specified. In the Netherlands it is common practice to obtain insight in the frequency of flooding or the occurrence of spills of wastewater into open water courses by applying a hydrodynamic model with a hydraulic load corresponding to a certain return period.

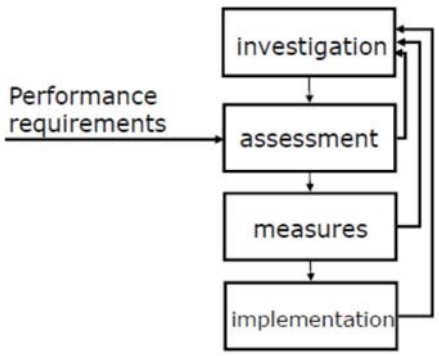


Figure 4: Integrated Sewer System Management Process (Nederlands Normalisatie Instituut, 2008)

The process scheme in Figure 4 shows the link between measures taken to adapt a sewer system based on the functioning of the system compared to the standard maintained. Since alterations in the sewer infrastructure are accompanied by high investment costs and a long lifespan it is important that the information source is reliable.

1.1 Problem

Hydrodynamic models are subject to uncertainties originating from various sources such as database errors, simplification of hydrological processes, numerical errors, failing sewer components,

incomplete description of processes, etc. (Korving, et al., 2002). These uncertainties influence the accuracy of the model.

The information obtained from a monitoring network has a higher level of accuracy compared to hydrodynamic models (Tait & ten Veldhuis, 2011), but is limited to the location, quantity and timespan measured. Moreover, monitoring campaigns alone fail to answer the *"what if"* question which on the short term is important to assess the effect of proposed alterations. On the long term, information from monitoring campaigns can be used as input for policy evaluation. A drawback is that although one can assess if goals are achieved, it is not possible to directly perceive if the policy maintained let to an optimal situation.

Uncertainties in the model results can be reduced when calibration of the model is applied. Calibration of a model results in a set of parameters that optimally fits the model to measured data. However, the portability of the obtained parameter values to other storm events is poor (Henckens, 2003). In other words, water level predictions for a storm event other than the storm event used to calibrate the model are less reliable. This is mainly due to two reasons; variation of parameters over time and an incomplete description of processes that influence the quantity measured. The latter can be exemplified using Figure 5; a certain process with output y is modelled by a first degree polynomial of the form ($y = bx+c$), while in fact the process is governed by an arbitrary second degree polynomial ($y = ax^2+bx+c$) influenced by a measurement error.

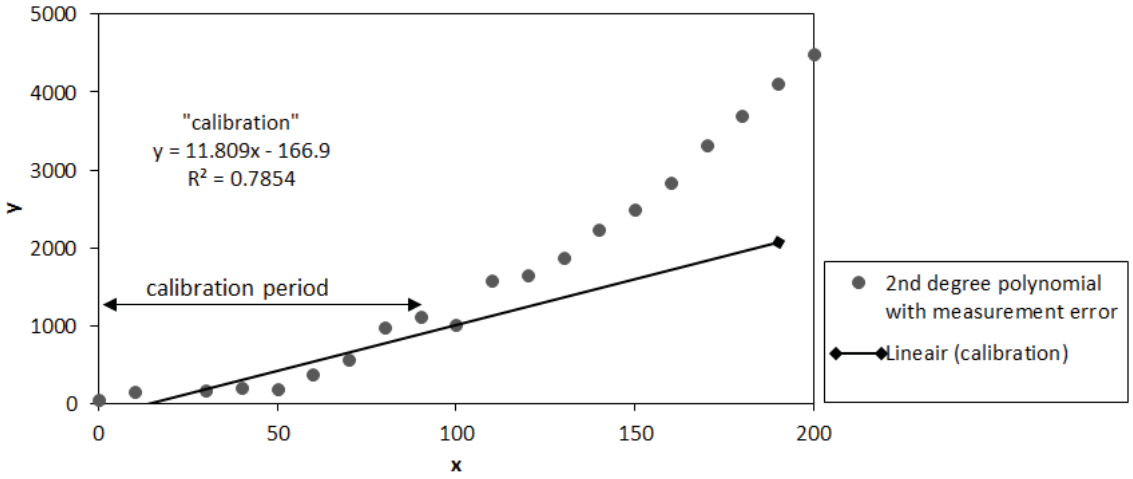


Figure 5 : modelling a second degree polynomial with first degree polynomials

The portability of the values obtained for b and c is limited, since these values incorporate a compensation for the missing term of the true polynomial due to an incomplete description of the process by the model. Therefore, predictions outside the calibration period are less reliable. This figure also shows that measurement errors do not only influence the portability, but also the model uncertainty within the calibration period.

An alternative is the use of multiple storm events in continuous time series for model calibration. However, it has been found that single event calibration results in a better match for a specific storm event (Henckens, et al., 2007). This phenomenon can again be exemplified by Figure 5; if the calibration period is extended, i.e. include more measurements in the calibration process, the new values for b and c will result in a deterioration of the match in the original calibration period.

1.2 Research aim

From the previous paragraph one can deduce that in their present form calibrated hydrodynamic models are unable to accurately reproduce measured water levels when using continuous time series.

The process of data assimilation is aimed at combining models and field observations. This concept has already been widely applied in combination with oceanic models and atmospheric models (Wang, et al., 2000). The principle of data assimilation is presented in Figure 6. The set of model parameter values is updated when new measurements are available. Therefore the deviation between the model output and the measured data is reduced over a longer period.

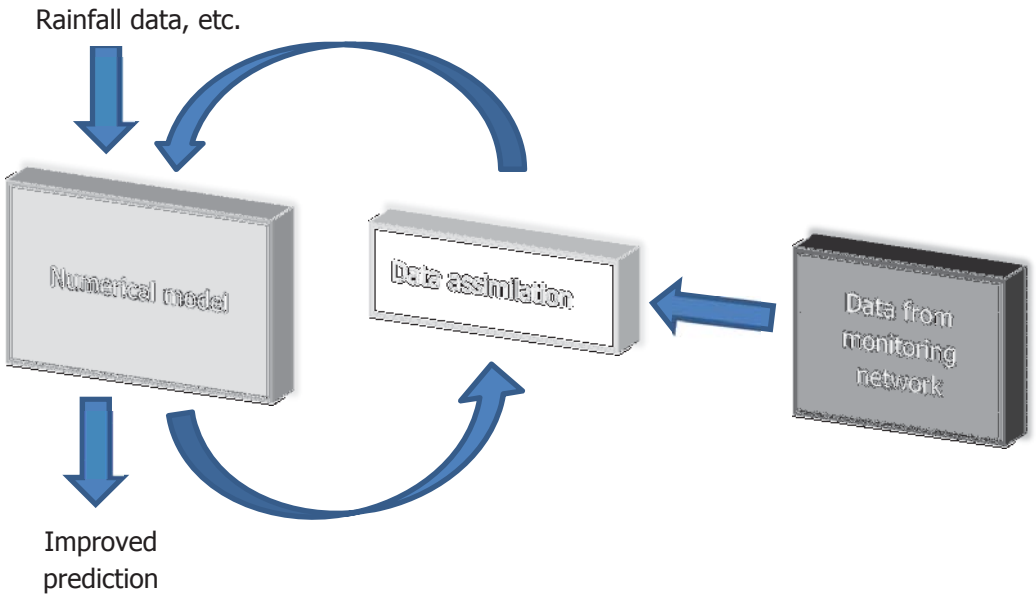


Figure 6: Principle of data assimilation

The principle aim of thesis is to:

Investigate whether data assimilation can be applied to models in urban drainage in order to simulate field observations for continuous time series.

Data assimilation requires information concerning the state of the system in order to correct the numerical model. Information on the relevant parameters is collected by a monitoring network. In this context, a monitoring network is defined as a set of sensors placed in either the sewer system or surface water system. A method will be elaborated to identify locations that are able to provide information on the relevant parameters. With respect to the measuring density in time, a sampling interval needs to be determined which is high enough to obtain sufficient information on the relevant processes.

From the main objective of this thesis the following research questions are derived:

- How can potential monitoring locations that are able to provide sufficient data for data assimilation be identified?

A methodology is elaborated that uses the results of a hydrodynamic model to find a set of locations most fit for collecting information on the relevant parameters.

– Can a hydrodynamic model be used to derive an upper boundary for the sampling interval of the monitoring network?

The effect of an increasing sampling interval on the information content is evaluated, and used to derive a measuring frequency for the case study.

– Is the applied data assimilation method feasible for application in the field of urban drainage?

Depending on the computational load needed to obtain reliable results, the applied data assimilation method may become unpractical for some purposes.

1.3 Possibilities for data assimilation

The concept of data assimilation can be applied to improve policy decisions (Shi & van Albada, 2007). Scenarios based on the results of data assimilated models can be used to evaluate the efficiency of policies. This creates the opportunity to obtain insight in the effects of investments in the water system. For instance the damage that is avoided during a certain storm event because investments in the sewer system have been made. This allows for optimal management of the system, where optimal can be referred to as minimizing the sum of yearly charges (Ven, van de, F.H.M., 2011).

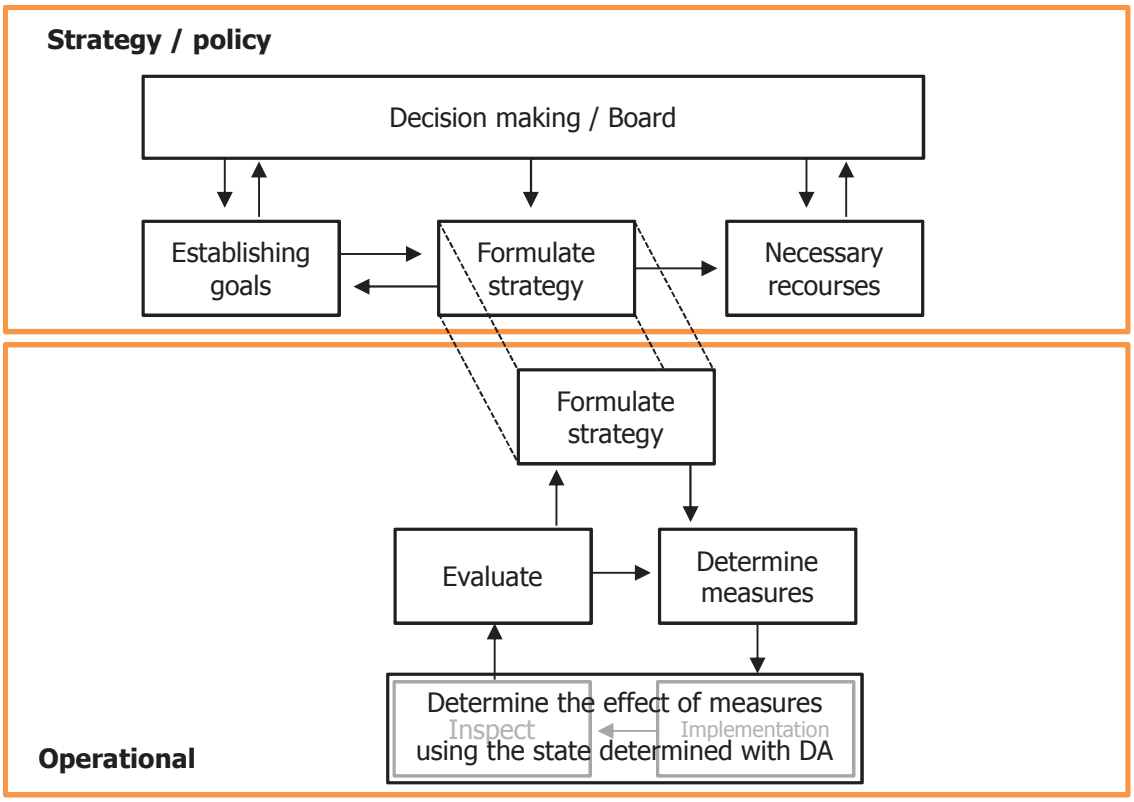


Figure 7: Management operations adapted from (Nederlands Normalisatie Instituut, 1994)

Figure 7 shows the location of data assimilated models in management operations as a source of information for the formulation of strategies instead of implementation and inspection. It should be noted that depending on the alterations suggested, modifications to the system may influence the derived state, therefore making the obtained prediction less reliable. When the suggested measures are implemented and a new state is derived, data assimilation can be applied to evaluate the effects of the changes as suggested in Figure 4.

1.4 Outline of the thesis

This thesis can be divided in to two parts. The first part deals with the design of a monitoring network to collect information on the relevant parameters. The second part is concerned with the application of data assimilation. Both parts are applied to a case study. The relations between the different chapters are presented in Figure 8. Up to now the relevance of the topic has been described along with the aim of this thesis.

The literature review in chapter 2 is considered to be a foundation for the current research in addition to indicating the significance within the framework. Moreover, gaps in literature are identified.

Chapter 3 introduces the case study and the characteristics with respect to the urban water cycle. This area is used as case study in the following chapters.

A method for the design of a monitoring network capable of providing sufficient data on the relevant parameters is described in chapter4. Besides optimising the information content, a de-correlation algorithm is introduced that allows for some overlap in the information collected. In combination with a genetic algorithm, this method is successfully applied to the case study area in chapter 6.

Theory on the data assimilation method applied in this thesis is discussed in Chapter 5, where an example compiled in Matlab® is presented. In Chapter 7 the implementation of Data assimilation for a commercially available hydrodynamic software package is discussed and applied to a simple model and the case study.

Finally, conclusions and recommendations based on the findings in these chapters are presented in Chapter 8.

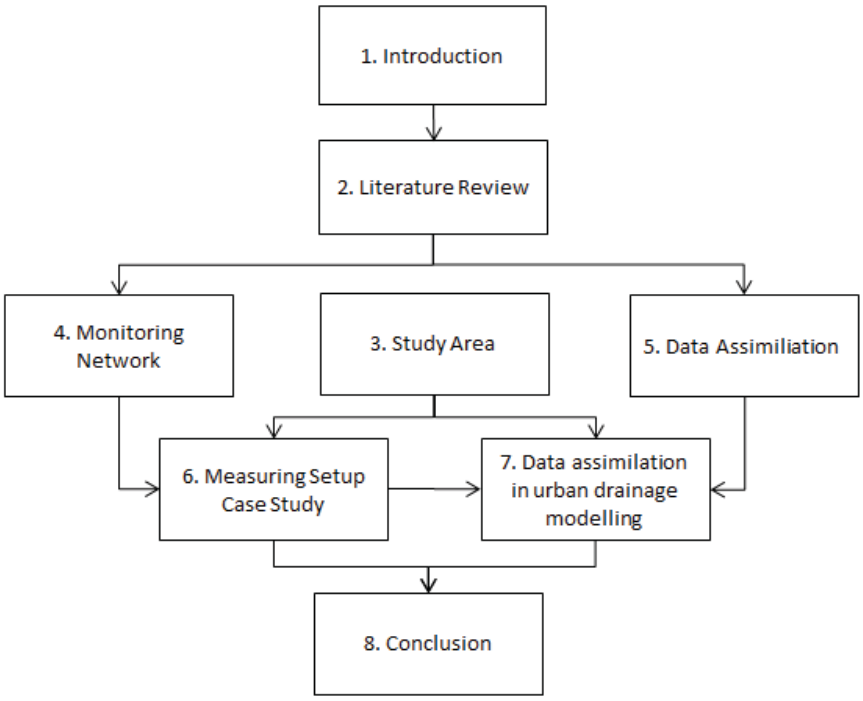


Figure 8: Thesis structure

2 Literature review

The articles included in this review provide a more in depth description of the topics introduced in Chapter 1. The significance of the calibration of models in urban drainage is covered. Hereafter a distinction is made between static calibration and dynamic calibration, where the added value of dynamic calibration for time series is demonstrated as well as identifying the gaps in literature. The information requirements calibration imposes on the design of a monitoring network is discussed in the last section of this chapter.

2.1 Significance of model calibration

The added value of a calibrated model is described by (Korving, et al., 2002). In this paper it is mentioned that investments in sewer rehabilitation are often made based on model predictions that are subject to uncertainty. Different types of uncertainties are identified and classified. Model parameter uncertainty is described in more detail and is analysed by applying Monte Carlo simulation on a reservoir model. The results for a particular case study show that this uncertainty can be quantified when a calibrated model is applied. (Clemens, et al., 2005) identify the main value of a calibrated model to be the quantified quality of the simulation per calibrated storm event and the parameter set produced. additional benefits can be found in the possibility to detect discrepancies in the structural database, obtain insight in system parameters such as weir coefficients, or to identify deficiencies in the model applied. Assessment of the quality of the calibration is stated by (Henckens, 2003) to be one of the main objectives of model calibration. In the case of Loenen in the Netherlands discrepancies were found during calibration that led to the findings of geometrical errors in the database.

2.2 Deficiencies of static calibration

Most of the research on the calibration of urban drainage models carried out up to this date is focussed on static calibration (e.g. (Clemens, 2001), (Di Pierro, et al., 2005) and (Kleidorfer, et al., 2009)). In this context, static refers to a procedure in which the parameter values obtained remain constant in time. For the case of Loenen in the Netherlands (Henckens, 2003) mentions that for different storm events different sets of parameter values are found. It is recognized that the obtained parameter sets cannot be directly compared since they have been derived with varying degrees of certainty. It should however be noted that the difference in parameter sets is influenced by several processes that are measured but are not incorporated in the model. Therefore water level predictions for a storm event other than the storm event used to derive the parameter values will be less reliable.

Subsequently (Henckens, et al., 2007) state that a parameter set obtained from several rain events will improve water level predictions. To this end continuous time series are used as input for a case study. Results show that the parameter set obtained by time series calibration deviates significantly from the parameter set obtained from the calibration using one of the storm events in the time series. It is found that single event calibration does provide a better match for a specific storm event. This is likely caused by the influence of hydrological processes and sediment transport on the long term. Therefore, static calibration is regarded to be unsuccessful in water level prediction for time series.

2.3 Dynamic calibration for time series

In contrast with static calibration, dynamic calibration results in a set of parameter values that is time dependent. (Rauch, et al., 2011) attempts to reduce the uncertainty in model predictions by proposing an algorithm that updates model parameters when new measured data is available. If the parameter value estimate distribution deviates too much from the previous distribution either the system has

changed, or the measurements are wrong. The former can be identified if multiple consecutive parameter value estimates show too much deviation, indicating the need for a new calibration. In the event of the latter, the deviation will only occur once. The added value of and update of the parameter values is seen in Figure 9, where an update results in a decrease in uncertainty.

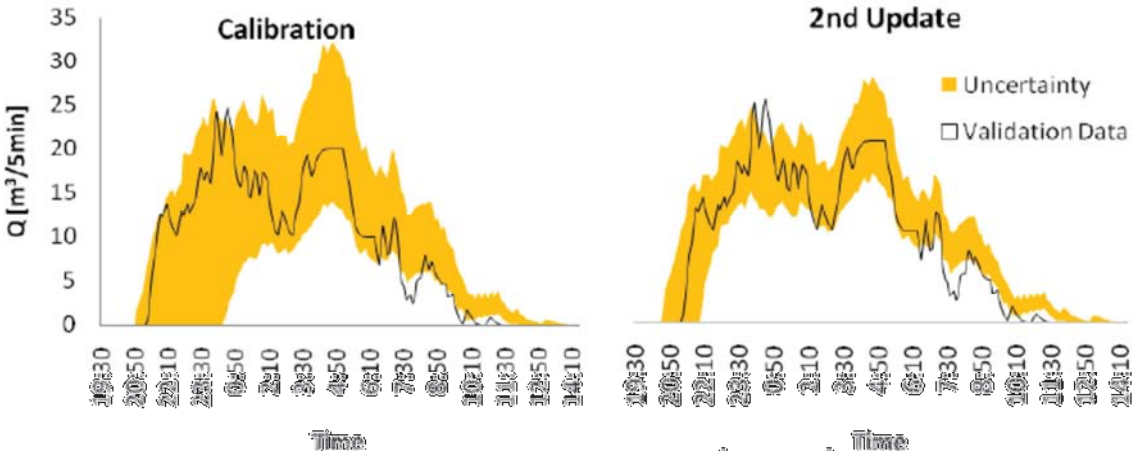


Figure 9: Uncertainty bandwidth in model predictions (20th and 80th percentile) and validation data after calibration on the left and after 2nd update on the right (Rauch, et al., 2011)

Moreover, this confirms the observation made by (Henckens, 2003) concerning the reliability of water level predictions for a storm event, other than the storm event used to derive the set of parameter values.

data assimilation addresses these uncertainties by updating the model parameter values when new field observations are available. Therefore deviations between the model and the measured data are minimized throughout a time series simulation. In other fields, data assimilation has already been widely applied. For ocean models, Data assimilation is recognized as the most powerful tool to improve consistency between the model and observations (Korres, et al., 2007). Although there is some experience with the application of data assimilation for looped networks for drinking water distribution (Kang & Lansey, 2009), no cases have been found concerning hydrodynamic models in urban drainage. Different data assimilation methods are discussed by (Hutton, et al., 2010). The authors report that the ensemble Kalman filter is less vulnerable to the non-linearity that is inherent to looped networks, compared to the Kalman filter. The functioning of the Kalman filter and ensemble Kalman filter is further elaborated in chapter 5.

2.4 Measurements for model calibration

(Ghil & Malanotte-Rizzoli, 1991) recognize a key problem for oceanographic assimilation that also applies to the assimilation of data in the field of urban drainage. How can one derive the state from one part of the system by using data derived from other parts of the system. This is dependent on the evolution of information through the flow. Therefore it is crucial that a monitoring network is designed in such a way that locations with a high information content are identified and monitored.

(Clemens, 2001) states that the design of a monitoring network and the calibration of a model are interrelated and should not be viewed as separate components. (Henckens & Clemens, 2004) describe a method that can be applied to design a monitoring network that meets the information requirement for calibration while minimizing investment costs. Locations that potentially provide the most information on a parameter are identified by calculating the influence that variation in parameters have on the model results. To prevent the same information from being collected multiple times,

correlation between sensors is punished by a de-correlation algorithm. However, (Henckens, et al., 2005) state that if correlation between sensors is punished, the monitoring network is more susceptible to information loss in case of sensor failure. Correlation between sensors can also be used for the cross-validation of data. Expansion of the de-correlation algorithm and theory on the information content is elaborated in chapter 4.

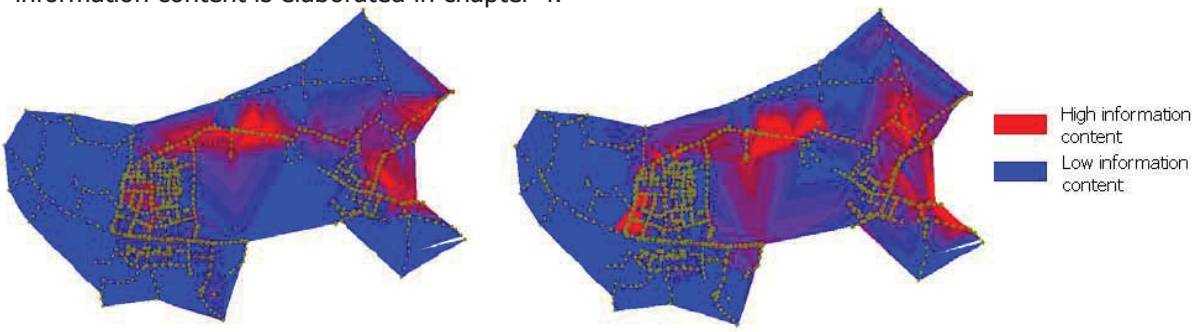


Figure 10: Total information content during two rainfall events (Henckens & Clemens, 2004)

The same authors found that the layout of the monitoring network obtained by applying the method described by (Henckens & Clemens, 2004) is very dependent on the storm event used, as seen in Figure 10. Hence, the use of multiple storm events is suggested. (Kleidorfer, et al., 2009) emphasize the relevance of the number of storm events and their characteristics instead of a fixed time period. The results presented suggest that if the wrong storm events are chosen, calibration will not be possible or an increase in the number of sensors is required.

Chapter Summary

- The match between simulated water levels and measured water levels obtained with single event calibration, cannot be achieved with time series calibration.
- Data assimilation addresses this problem by updating the set of parameter values when new field observations are available. However, applications in the field of urban drainage are lacking.
- The process of calibration imposes requirements on the monitoring data collected. Sufficient information on the relevant parameters is needed, while some overlap in the information collected is needed to increase the overall robustness of the monitoring network.

3 Case study characteristics

In the following chapters the sewer system of Delft's city centre is used as a case study for the application of data assimilation. A monitoring network is also designed for this system, in order to provide the necessary observations. This chapter introduces the case and the characteristics of the system.

Delft was founded over 750 years ago, and experienced a rapid expansion early on (Municipality of Delft, 2012). Boundaries of the city centre are presented in Figure 11. The same illustration shows the different canals that run through the city centre.

Figure 11: Aerial photo of Delft's city centre (Municipality of Delft, 2012)

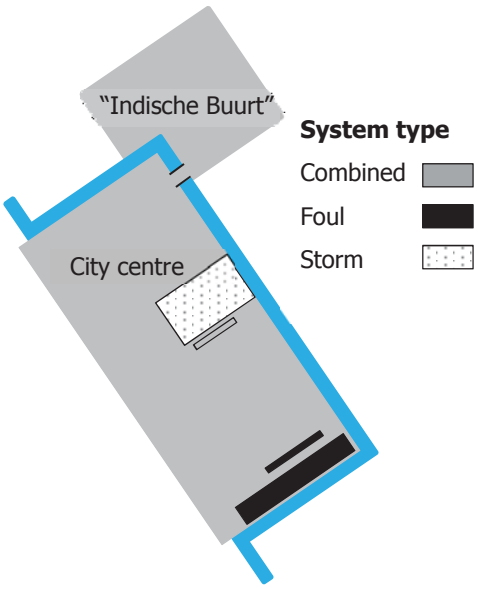


Figure 12: schematic representation of the relevant sewer districts

Like most urban areas in the Netherlands, Delft's centre has a combined sewer system. This entails that sanitary sewage and stormwater runoff is collected in a single system and transported to the wastewater treatment plant. During heavy rainfall the drainage capacity of the sewer system is insufficient and a combined sewer overflow (CSO) to the surface water will occur. As long as the surface water level is below the crest of the overflow weir the sewer system can freely overflow to the surface water.

The sewer system contains three subsystems; two small foul sewer systems and one combined. The storage of the main system below the lowest sewer overflow construction is 8.4 mm, taking into account 59.07 ha of connected surface area. This is considered to be a normal amount of storage in the Dutch situation. There is also a small storm sewer system located in the area. Characteristics of the main sewer system are found in Table 1.

Table 1: Characteristics of the Delft city centre sewer system

Urban drainage system characteristics	Delft city centre	
Number of inhabitants	8,590	
Storage volume	4962 m³	(= 8.4 mm)
Contributing area	59.07 ha	
Pumping capacity	838.8 m³/h	(= 1.42 mm/h)
Dry weather flow	103 m³/h	(= 0.17 mm/h)
Number of CSO structures	48	
lowest - highest crest level CSO structures	0.43 m-NAP	0.15 m-NAP

A connection is made close to the pumping station with the "Indische Buurt" sewer system as shown in Figure 12. Because the Indische Buurt area has lower ground levels, a height-adjustable weir is constructed just before the pumping station, preventing a large flow from the city centre system to cause flooding in the Indische Buurt.

Figure 13: Canal in the city centre of Delft (Dijk, van, Z., 2005)

One of the canals is shown in Figure 13, what is striking is the small freeboard. The small freeboard has the following implications for the sewerage system:

- A small increase in the surface water level can cause inflow of surface water in the sewer system through sewer overflow constructions,
- Only a small increase in energy levels in the sewer system can be tolerated before pluvial flooding occurs

The latter is addressed by constructing a large amount of sewer overflow constructions, namely 49. However, model calculations still estimate pluvial flooding to occur locally more frequent than once every year when the Dutch design storms are applied. The crest level of the lowest CSO structure is -0.43 m NAP[1], which is equal to the target water level of the surface water system. The former is dealt

[1] Normal Amsterdam Water Level is abbreviated to NAP

with by installing a set of adjustable weirs (see Figure 15) in the canals in order to create a temporarily isolated water system with its own target water level. The boundaries of this system are presented in Figure 14. When heavy rainfall is expected, the switch on level of the pumping station is set 0.20 meters lower than the normal target water level of -0.43 m NAP meter to create extra storage.

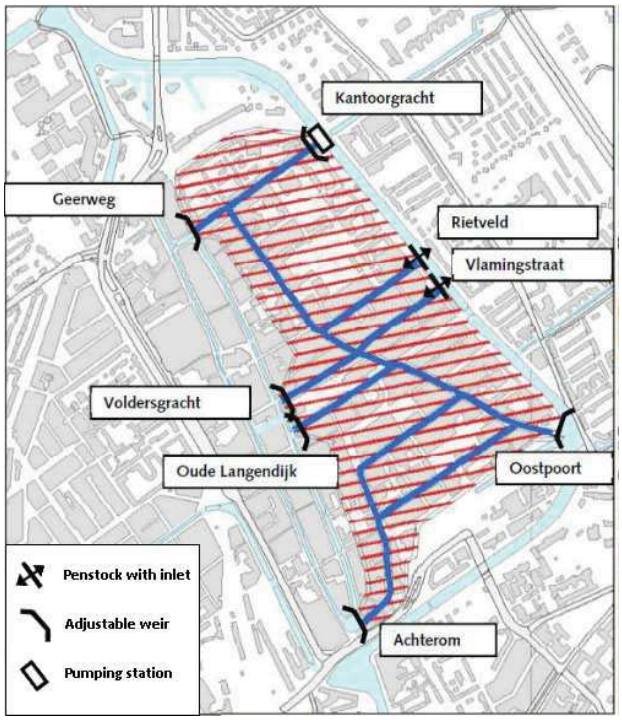


Figure 14: Overview of the engineering works used to create the isolated water system

Figure 15: Adjustable weir in closed position

Creating an isolated water system with a lower water level does not solve the problem completely, due to the fact that sewer overflow constructions are located on both water systems. Therefore the sewer system can act as a connection between the water system along the outer border and the isolated water system created by the adjustable weirs, provided that the water level is above the crest level of lowest CSO structure on both sides. This scenario is schematized in Figure 16.

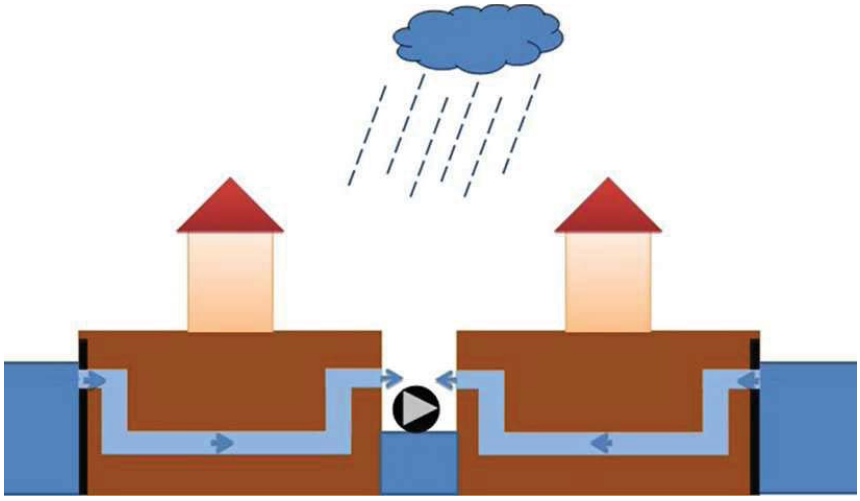


Figure 16: Scenario where the sewer system transports water from one surface water body to another

As a result of this interaction, the available storage in sewer system and the overall drainage capacity will decrease. Therefore pluvial flooding can be expected to occur more frequently. Since the CSO discharges will also increase, not only the volume of raw sewage increases, but this volume will also be spread over fewer sites, causing a progressive deterioration of the water quality locally. The interaction between both systems is also amplified by the characteristics of the Delfland Basin, which covers an area urbanised to a large extent. This results in a higher peak discharge in case of a storm event, compared to a rural area.

Chapter Summary

- The Delft city centre drainage system is susceptive to pluvial flooding due to the limited freeboard.
- Since the crest levels of the sewer overflow weirs are low compared to the target surface water level, interaction between the surface water system and the sewer system is likely to occur during storm events.
- A series of adjustable weirs have been constructed in the surface water system to lower the target water level during storm conditions.

4 Monitoring network

In order to gain information on relevant processes in the field of urban drainage, data originating from various sources is collected. Although hydrodynamic models are widely applied to collect data, other sources such as sewer inspections, flooding records from the fire brigade, monitoring networks, call centres, pump operating hours, media etc. are also identified. The aim of this chapter is to describe a methodology that can be used to design a monitoring network that is able to provide data which can be assimilated by the hydrodynamic model in order to reduce the uncertainty in model results.

The first paragraph introduces the concept of establishing the information requirement in order to determine what data should be collected. This line of thought is incorporated in the second paragraph where a method is described that can be used to design a monitoring network. This method is applied to a case study is presented in chapter 6.

4.1 Information requirement

According to (Lohuizen, van, C.W.W., 1986) the translation from data to an actual decision can be described by the scheme presented in Figure 17. This scheme shows the steps taken to convert data into a decision.

Figure 17:The knowledge household (Lohuizen, van, C.W.W., 1986)

Data can be defined as uninterpreted characters, signals, patterns that have no direct meaning for the system under observation (Aanmondt & Nygard, 1995). A selection of the available data is made and this information is analysed to increase the knowledge of the system. Interpretation of this knowledge to form an image of the current situation and comparison with the desired situation yields the necessary understanding to make a decision concerning the system.

This scheme can also be used from top to bottom. If the decisions a stakeholder ought to make are known, these needs can be translated to data required to make this decisions. This is an important link in designing an monitoring network for the collection of data. (Langeveld, et al., 2004) state that for each process and water quality parameter studied the optimal configuration of a monitoring network can be different. This implies that if it is not well known what information is needed to make decisions, the optimum monitoring network to provide the necessary data is not likely to be found.

This line of thought can be extended to the calibration of models. If a calibrated model is regarded as an information source to improve the knowledge of the system, the data collected first needs to be converted into information that is used by the model to accurately simulate the system of interest.

Therefore, in this chapter the information requirement for model calibration is determined. Relevant model parameters are identified by quantifying the effect of a change in parameters on the model results. Subsequently, a monitoring network is designed that maximises the information content on these parameters.

It should be noted that a monitoring network often proves to be non-static. Evaluation of the obtained information may lead to new or redefined information need (UN/ECE Task Force on Monitoring & Assessment, 2000). In Figure 18 the monitoring cycle is presented. This figure shows that a new information need can result in an update of the monitoring program.

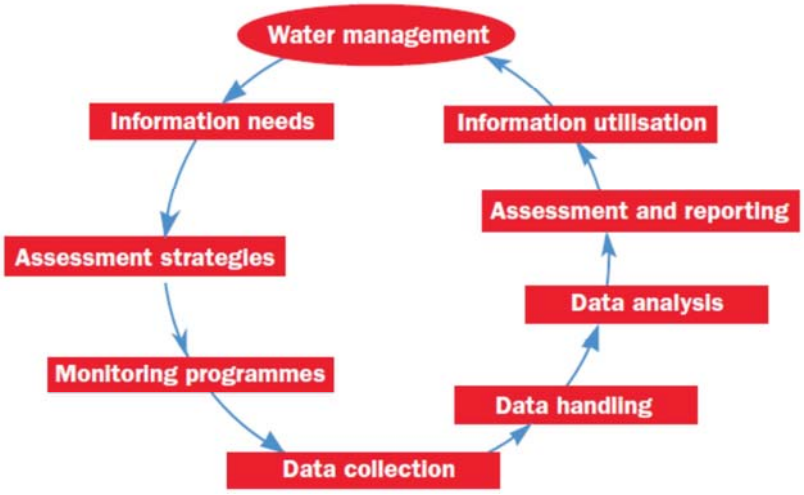


Figure 18: Monitoring cycle (UN/ECE Task Force on Monitoring & Assessment, 2000)

4.2 Design of a monitoring network

Several choices have to be made in the process of designing a monitoring network. This can be approached as a constraint optimisation problem, where the information content should be optimized taken into account the available budget (Henckens & Clemens, 2004). The information content can be defined as the amount of linear independent pieces of information that are contained in a set of observations from a monitoring network (Johnson, 2003). It should be noted that the information content is not only influenced by the apparatus chosen, but also by the location and the time interval between measurements. Therefore, if the information requirement is known, the following questions need answering:

- what type of measurement is required? section 4.2.1
- what is the measuring accuracy and frequency needed? section 4.2.2
- Where should the sensors be placed? section 4.2.3

4.2.1 Methods for water quantity measurements

In order to answer the first question, the water quantity to be measured is identified. This decision is made based on the general suitability of the type of measuring device with respect to the system under observations, and the accuracy of the measuring method.

Not taking into account the rainfall measurements, water levels and discharges are the quantities commonly measured in urban drainage. Depending on the measuring principle water levels can be

measured with a high accuracy. Pressure head measurement has a measuring accuracy of approximately 2mm (Rioned, 2003). Methods using ultra sound to measure water levels are able to achieve a similar accuracy, but have a smaller measuring range due to the fact that sufficient distance between the sensor and the water level is required (Rioned, 2009).

For discharge measurements different principles can be applied. Compared to water level measurements, measuring accuracies reached are worse while the installations are more expensive (Clemens, 2001). Therefore the monitoring network designed will be based on data obtained from water level measurements. This will not only result in measurements with a higher accuracy, but also the possibility to have more monitoring locations with the same budget.

4.2.2 Measuring frequency and accuracy

This section is aimed at defining a measuring frequency based on the accuracy of the apparatus chosen. Noise originating from the measuring inaccuracy corrupts the signal, therefore making it more difficult to reconstruct the original process. Beyond a certain measuring frequency, the extra information obtained will be dominated by this noise and will not result in an increase in information.

When data produced by a hydrodynamic model is used to obtain an estimate for the sampling frequency this noise term needs to be added, since the model produces the 'original signal' as seen in Equation (4-1). The noise term is modelled as a series of random independent samples taken from a normal distribution with a zero mean μ. The standard deviation σ describes how concentrated the distribution of samples is around the mean, and is inherent to the accuracy of the measuring device.

$$f(t) = h(t) + \varepsilon(t) \qquad (4\text{-}1)$$

Where:

f(t) = signal that is sampled by a measuring device
h(t) = real signal
Ɛ(t) = noise originating from measuring inaccuracy

An upper boundary for the sampling frequency can be determined by analysing the signal in the frequency domain as described by (Clemens, 2001). In the frequency domain information on the strength of a signal within a certain frequency range is made visible. This is demonstrated in Figure 19, where a sinusoid of 120 Hz and a sinusoid of 50 Hz are plotted. the different frequency components are not distinguishable in the time domain but are in the frequency domain.

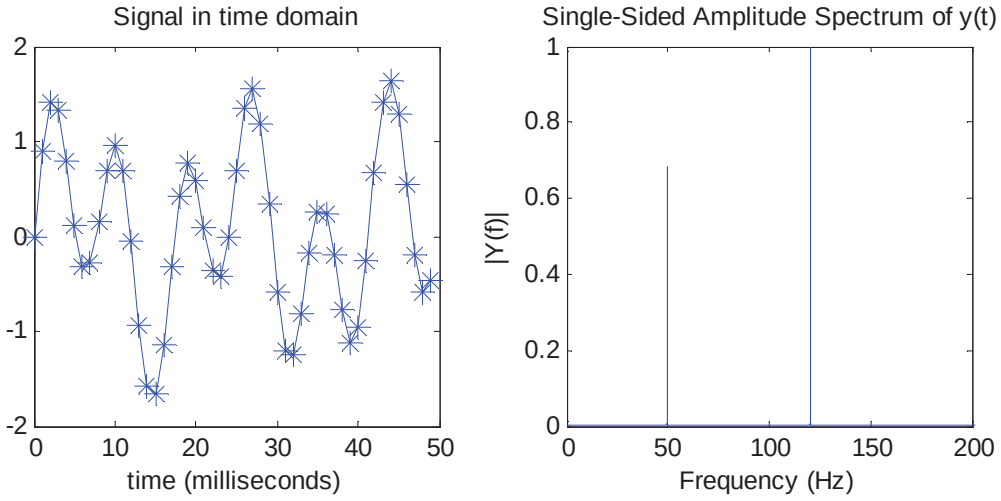


Figure 19: example of a 50 and 120 Hz sinusoid in the time domain and frequency domain (Mathworks, 2012)

Conversion to the frequency domain is achieved by a Fourier transform. This entails the decomposition of a signal into a series of trigonometric functions (Butz, 2006). This conversion does not change the information content, but presents the same information in a different way. The Fourier transform for a discrete signal with a finite duration is described by Equation (4-2) and the inverse transform by Equation (4-3), see e.g. (Arfken, 1985).

$$F_n = \sum_{k=0}^{N-1} f_k e^{\frac{-2\pi \cdot i \cdot n \cdot k}{N}} \qquad (4\text{-}2)$$

$$f_k = \frac{1}{N} \sum_{n=0}^{N-1} F_n e^{\frac{2\pi \cdot i \cdot n \cdot k}{N}} \qquad (4\text{-}3)$$

Where:

N = number of observations
i = imaginary unit
k = 0, 1, 2, ..., …, $N - 1$

The strength of variations as a function of the frequency can be seen by exploring the Power Spectral Density function (PSD). For the transform of a series of observations $f(t)$ to $F(\omega)$ the PSD function is defined by Equation (4-4).

$$F_{PSD}(\omega) = \sqrt{F(\omega) \cdot \overline{F}(\omega)} \qquad (4\text{-}4)$$

Where:

$F_{PSD}(\omega)$ = PSD function
$F(\omega)$ = transform of a series of observations
$\overline{F}(\omega)$ = complex conjugate of $F(\omega)$

A theoretical example of a PSD function is presented by the dash-dot lines in Figure 20. The original signal is composed of a function that increases linearly in time and the noise term is Gaussian white noise with a standard deviation of 0.05.

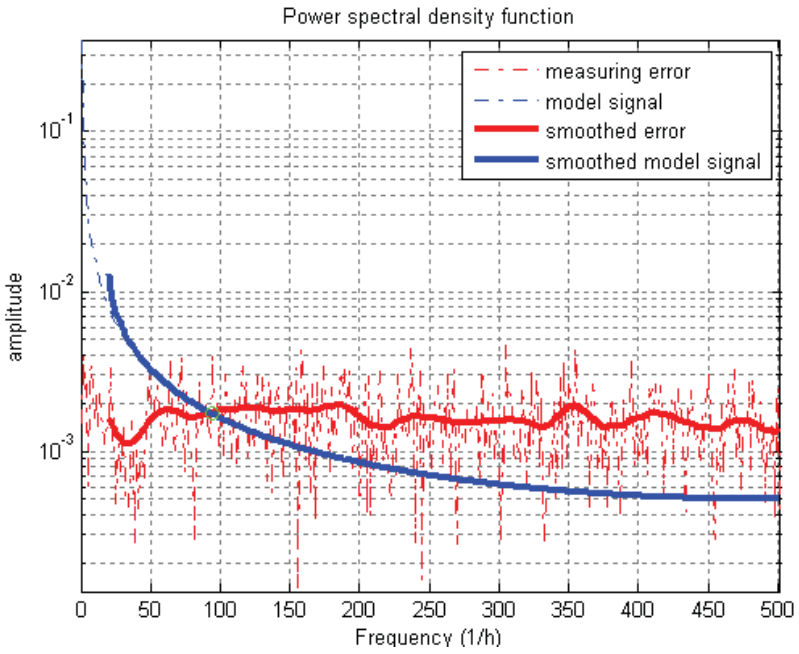


Figure 20: Example of a power spectral density function (PSD)

The smoothed lines are generated to remove the small scale variation which makes it more difficult to compare the two power density functions. This smoothing technique is proposed by (Hamilton, 1994) and calculates the weighted average of a number of observations around the point of interest (see Equation (4-5)). The number of observations taken into account is influenced by *h*, and the weights assigned to the observations increases when closer to the point of interest. The total sum of the weights equals unity. Values chosen for *h* are considered subjective. Too small values will not remove the variation, while large values will introduce some bias (Hamilton, 1994).

$$\hat{F}_{PSD}\left(\omega_j\right) = \sum_{m=-h}^{h}\left[\frac{h+1-|m|}{(h+1)^2}\right]\cdot F_{PSD}\left(\omega_{j+m}\right) \qquad (4\text{-}5)$$

Where:

$\hat{F}_{PSD}\left(\omega_j\right)$ = weighted average value for observation *j*

h = parameter for the amount observations taken into account ($2\cdot h+1$)

For example, when using *h = 1* Equation (4-5) becomes:

$$\hat{F}_{PSD}\left(\omega_j\right) = \frac{1}{4}F_{PSD}\left(\omega_{j-1}\right)+\frac{1}{2}F_{PSD}\left(\omega_j\right)+\frac{1}{4}F_{PSD}\left(\omega_{j+1}\right)$$

As seen in Figure 20, beyond a certain frequency the signal is dominated by noise originating from measuring inaccuracy and therefore will not result in an increased information content on the relevant process. Based on this observation, the maximum sampling frequency is defined by:

$$\hat{F}_{PSD}\left(\omega\right) \approx \hat{E}_{PSD}\left(\omega\right) \qquad (4\text{-}6)$$

Where:

$\hat{F}_{PSD}\left(\omega\right)$ = power spectral density function of the original signal

$\hat{E}_{PSD}\left(\omega\right)$ = power spectral density function of the measuring inaccuracy

According to Shannons theorem a signal can be reconstructed exactly from discrete samples when the sampling frequency is at least twice the original signal frequency (Nooyen & van Overloop, 2008). This is translated into Equation (4-7).

$$\omega_F < \frac{1}{2}\omega_s \qquad (4\text{-}7)$$

Where:

ω_F = frequency of the original signal

ω_s = sampling frequency

4.2.3 Spatial distribution of the monitoring locations

The process of designing a monitoring network is aimed at identifying locations that provide the most information on the process of interest. For model calibration this entails the gathering of information on relevant model parameters. Reasons to omit certain locations on forehand are discussed. The goal of a singular value decomposition of the Jacobian matrix is twofold; to determine the model parameters best suited for calibration, and to find the set of monitoring locations with the largest

information content with respect to these parameters. Subsequently a de-correlation algorithm is elaborated to control the overlap of information.

4.2.3.1 Excluding locations

It is not possible, nor reasonable to measure every possible location (Kleidorfer, et al., 2009). This is due to financial constraints, the amount of data that needs processing and various practical considerations. Some locations need be omitted based on accessibility. This can concern for instance manholes on private terrain, main traffic routes or other inaccessible locations (see Figure 21). It should be noted that the manholes also need to be accessible after installation for maintenance and repairs. Other manholes should be excluded based on hydraulic properties (Henckens, et al., 2005). This includes manholes with pipes attached that have a invert level above the manhole bottom, which causes unwanted turbulence. Measuring devices placed behind special structures (weirs, sluices etc.) can be subjected to a higher gas concentration which can influence the accuracy of the apparatus applied. These locations are removed later in the process to ensure that the maximum information about the system is obtained.

Figure 21: difficult accessible manhole

4.2.3.2 Sensitivity analysis

At a certain point, adding more monitoring locations has a negligible effect on the total information content for the parameters of interest. Furthermore not all locations provide the same amount of information on the same parameters. This means that a set of locations needs to be identified that provides sufficient information on all the relevant model parameters. The information content is determined by analysing the Jacobian matrix. This matrix shows the sensitivity of the water level to a change in a certain parameter. If a parameter value is varied and this does not result in a change in water level, that particular location at a certain time does not provide information on the parameter in question. The Jacobian matrix is computed using a finite difference approximation as seen in Equation (4-8). For this approximation, the model is run $n + 1$ times. One run with the original parameter set, and n runs where one of the parameter values is varied each time. The variation in the parameter values is typically 5% (Langeveld, et al., 2004).

$$\underline{\underline{J}} = \frac{\partial \underline{\underline{h_t}}}{\partial \underline{p_i}} \approx \frac{\Delta \underline{\underline{h_t}}}{\Delta \underline{p_i}} \qquad (4\text{-}8)$$

Where:

h_t = the water level at a certain location at time t obtained from a model

p_i = model parameter no. i

In matrix form for an arbitrary location x, Equation (4-8) becomes:

$$\underline{\underline{J}}_x = \begin{bmatrix} \frac{\partial h_{t=0}}{\partial p_1} & \cdots & \frac{\partial h_{t=0}}{\partial p_n} \\ \vdots & \cdots & \vdots \\ \vdots & \cdots & \vdots \\ \vdots & \cdots & \vdots \\ \frac{\partial h_{t=t\max}}{\partial p_1} & \cdots & \frac{\partial h_{t=t\max}}{\partial p_n} \end{bmatrix} \qquad (4\text{-}9)$$

Where the columns of the Jacobian show the sensitivity for a given parameter at different observations, and the rows express the sensitivity of one observation for the different parameters. The time interval between different discrete observations needs to be equal to the sampling interval of the sensor chosen.

4.2.3.3 Singular value decomposition

Although the Jacobian matrix provides direct information on the effect of a parameter value change on the water level at a certain time step, extra information originating from the change of the water level over time is not made visible. This information can be extracted by computing the singular value decomposition (SVD) of the Jacobian matrix. If the Jacobian matrix is denoted by $\underline{\underline{J}}$ than the SVD of this matrix is described by Equation (4-10).

$$\underline{\underline{J}} = \underline{\underline{U}}\,\underline{\underline{\Sigma}}\,\underline{\underline{V}}^T \qquad (4\text{-}10)$$

Where:

$\underline{\underline{J}}$ = *m x n* Jacobian matrix

$\underline{\underline{U}}$ = *n x n* matrix containing the left singular vectors of *J*

$\underline{\underline{\Sigma}}$ = *n x m* matrix containing the singular values of *J* as diagonal entries

$\underline{\underline{V}}$ = *m x m* matrix containing the right singular vectors of *J*

The singular values of $\underline{\underline{J}}$ are the square roots of the eigenvalues of $\underline{\underline{J}}^T\,\underline{\underline{J}}$ and define to what extend $\underline{\underline{J}}$ stretches or shrinks certain vectors, and can be used to find the dimensions along which data shows the largest variation. The direction in which the vector is stretched is defined by the eigenvectors of $\underline{\underline{J}}^T\,\underline{\underline{J}}$ that form the columns of $\underline{\underline{V}}$. An example for a simple case is shown in Figure 22, where the singular values are denoted by λ_i in a descending order, and the corresponding eigenvector by $\underline{v}$. It can be seen that the largest singular value represents the maximum stretching, while the corresponding eigenvector identifies the direction.

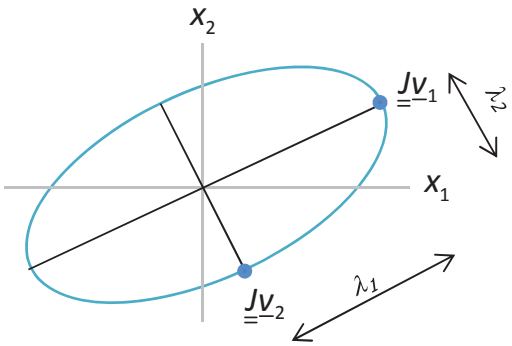


Figure 22: Schematic representation of singular values and eigenvectors for a ellipse (Lay, 2006)

If the Jacobian matrix is rank deficit, one or multiple singular values are zero. Although SVD is very reliable with respect to numerical errors (Kalman, 1996), round off errors can result in extremely small nonzero singular values. Therefore small singular values are assumed to be zero in this thesis. The effect of a singular value that is (almost) zero on the information content for a certain parameter is demonstrated using Equation (4-11) as stated by (Clemens, 2001) and (Olsthoorn, 1998).

$$\left.\begin{array}{l} \underline{\underline{J}} = \dfrac{d\underline{h}}{d\underline{p}} \rightarrow d\underline{h} = \underline{\underline{J}} \cdot d\underline{p} \\ \underline{\underline{J}} = \underline{\underline{U}}\underline{\underline{\Sigma}}\underline{\underline{V}}^T \end{array}\right\} d\underline{h} = \underline{\underline{U}}\underline{\underline{\Sigma}}\underline{\underline{V}}^T \cdot d\underline{p} \qquad (4\text{-}11)$$

$d\underline{h}$, $\underline{\underline{U}}$ and $\underline{\underline{V}}$ can be considered bounded, while $d\underline{p}$ can take on any value. So if the singular value approaches zero, a change in water level $d\underline{h}$ can only be observed for values of $d\underline{p}$ approaching infinity. From this, one can conclude that the information content for a particular parameter is small when the corresponding singular value is (nearly) zero. Throughout this thesis, parameters corresponding to (nearly) zero singular values are denoted as being unidentifiable. Identifiability problems may have two causes according to (Speed & Ahlfeld, 1996) and (Seber & Wild, 1989); data dependent causes and model dependent causes. The former is influenced by the measuring accuracy, sampling interval and the locations being monitored. The latter is inherent to the model applied.

With respect to the design of a monitoring network, singular values are used to identify model parameters eligible for optimisation and to judge the information potential monitoring locations can provide on these parameters. The former is achieved by removing parameters that correspond to (nearly) zero singular values from the parameter set used for optimisation. The latter is further elaborated in section 4.2.3.4.

The right singular vectors are used to link the parameters to the singular values. When a particular singular vector has more than one non-zero entries, the corresponding parameters both influence that particular singular value. this means that the individual parameters cannot be identified separately.

4.2.3.4 Optimisation of the information content

In addition to using singular values to calculate the information content of the system, singular values can also be used to calculate the information content for a particular location. To this end a singular value decomposition of the part of the Jacobian referring to the location in question is calculated. The relative information content for this particular location can be described by Equation (4-12) (Henckens, et al., 2005). This equation divides the set of singular values obtained from one location by the singular values calculated for the entire system in order to normalize the information content.

$$Ic_x = \frac{\sum_{i=1}^{P_{tot}} \frac{Ic_{xi}}{Ic_{ni}}}{P_{tot}} \qquad (4\text{-}12)$$

Where:

Ic_x = relative information content of location x

Ic_{xi} = information content of location x for parameter i

Ic_{ni} = information content of the system for parameter i

P_{tot} = number of parameters

This equation provides a basis for the design of a monitoring network that can be used for model calibration, since parameters that have a larger influence on the water level are favoured. If the goal of monitoring is to measure a specific parameter, but the water level is insensitive to variations of this parameter a different approach is needed. This will likely result in a more expensive monitoring network due to a large amount of sensors (Henckens & Clemens, 2004).

It should be noted that the storm event chosen for this sensitivity analysis has a large influence on the information content. This is logical since it is impossible to obtain information on a weir coefficient if the storm event used does not result in a CSO. This can be dealt with by using multiple storm events for the design. The amount of storm events used is limited by the maximum allowable size of the Jacobian (Henckens, et al., 2005).

Another point of concern is the possible correlation between different sensors. If two locations show a high correlation, the relative information content of the individual locations might be high but this will not yield a high combined information content since the information provided has a large overlap. This can be addressed by punishing one of the locations, therefore making it less likely that both locations are monitored.

The following algorithm is proposed by (Henckens, et al., 2005) for the case where three sensors are present. Since a small correlation is not considered a problem, an overlap allowance can be specified.

1. The locations are ordered in descending order with respect to the information content
2. All except the first location are de-correlated by multiplying the information content with a weight ≤ 1 depending on the correlation between the two locations.
3. The first sensor is taken out of the set, and step 1 is repeated until there are no more sensors left.

For the case where three sensors are present:

$$\begin{aligned} Ic_1 &= Ic_1 \\ Ic_2 &= Ic_2 \cdot W_{21} \\ Ic_3 &= Ic_3 \cdot W_{31} \cdot W_{32} \end{aligned} \qquad (4\text{-}13)$$

Where:

Ic_x = relative information content of location x

W_{xy} = weight based on the normalized cross-correlation between sensor x and y

This algorithm is largely based on the principle that a high correlation between locations is punished. The punishment is used to diminish the information content and therefore reducing the chance that this location is chosen to be monitored. The weight factor can be calculated using Equation (4-14) which is a slightly modified version of the equation mentioned by (Henckens & Clemens, 2004) .

$$W_{xy} = \begin{cases} 1 & \text{if } C_{xy} \le O \\ \dfrac{1-\left|C_{xy}\right|}{1-O} & \text{if } C_{xy} > O \end{cases} \qquad (4\text{-}14)$$

Where:

W_{xy} = weight based on the normalized cross-correlation between sensor x and y

O = the overlap allowance (0 – 1)

C_{xy} = the normalized cross correlation between location x and y

The overlap allowance reduces this punishment and provides some room for overlap. An example of this linear weight function is shown in Figure 23. For a correlation smaller than the overlap allowance the location is not punished, while for a correlation of unity (correlation with itself) the information content is reduced to zero.

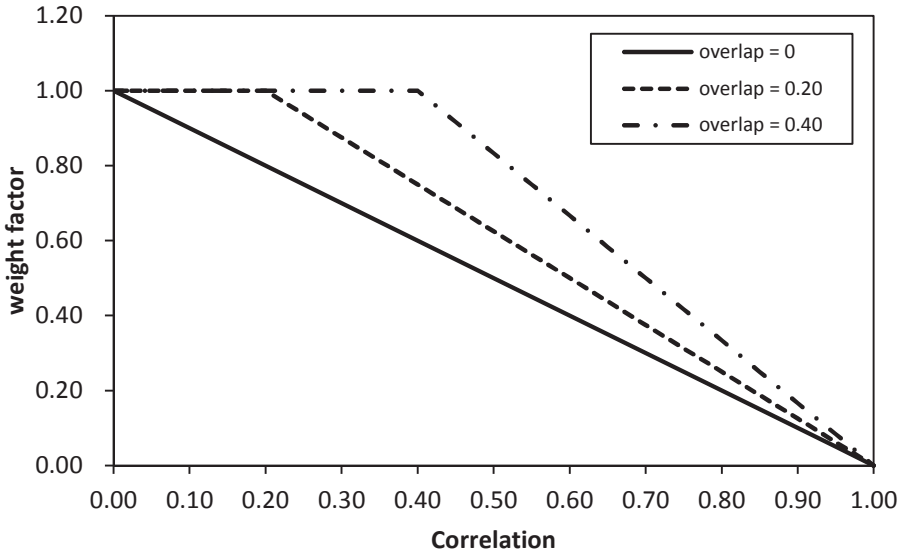


Figure 23: weight factor for different overlap values

Applying this algorithm results in a set of locations where information from one particular location cannot be obtained (well) from another location in the set. This has the advantage that the same information is not collected multiple times. Disadvantages can be found in the fact that if there is no overlap and a sensor breaks down, potential valuable information is lost. Correlation between sensors can also be used for the purpose of validation, to identify faulty measurements. (Harder, 2010) states that for the purpose of data validation each sensor should have at least one well correlated sensor. He defines a well correlated sensor to be $C_{xy} > 0.8$, while for some sensors good results are obtained for $C_{xy} > 0.6$.

In order to meet these requirements, Equations (4-13) and (4-14) need to be redefined. Instead of punishing all other locations with a weight proportional to the correlation, **one** location with a correlation closest to a predefined value will be weighted using a different equation. In order to promote the sensor with a correlation near the predefined value and punish the others a second de-correlation function is proposed. Several weight functions which foster points with a high correlation and neglect other points are eligible. In this thesis a set of linear functions is chosen due to its simplicity (see Equation (4-15)). Several test runs have also been made in this thesis with Gaussian weight functions as mentioned by (Brubaker, 2006), where the standard deviation is used to assess the correlation This resulted in the same set of monitoring locations.

$$W_{xy} = \begin{cases} 0 & \text{if } C_{xy} < \alpha \\ \frac{1}{\beta - \alpha}\left(C_{xy} - \alpha\right) & \text{if } \alpha \le C_{xy} < \beta \\ 1 & \text{if } \beta \le C_{xy} < \gamma \\ \frac{1}{\gamma - 1}\left(C_{xy} - 1\right) & \text{if } \gamma \le C_{xy} \le 1 \end{cases} \qquad (4\text{-}15)$$

Where:

W_{xy} = weight based on the normalized cross-correlation between sensor *x* and *y*
C_{xy} = normalized correlation between sensor *x* and *y*
α = lower boundary for the cross correlation which values below get a weight factor 0 allocated (0-1)
β = minimum cross correlation where the maximum weight factor is assigned (0-1)
γ = maximum cross correlation where the maximum weight factor is assigned (0-1)

An example of this function is seen in Figure 24. A lower boundary of 0.50 is chosen, which results in a weight factor of ⅓ at the lower boundary of correlation reported by (Harder, 2010). The weight factor is unity between 0.7 and 0.9 and decreases to be zero at a correlation 1. The weight factor is required to be zero at a correlation of unity to prevent a specific location from correlating with itself.

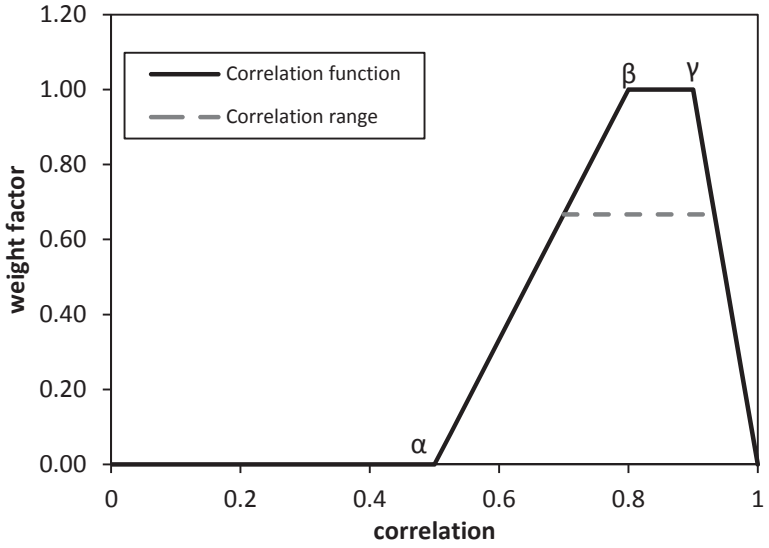


Figure 24: weight factor for different values of the standard deviation and μ = 0.8

A drawback of this approach is that *situation 1* as schematically shown in Figure 25 can occur, while *situation 2* is preferred. *Situation 1* shows that location A has a good correlation with location B, but location B has a slightly better correlation with location C etc.. This will severally limit the diversity of the total information content. *Situation 2* on the other hand shows a network where each sensor is well correlated with one other sensor.

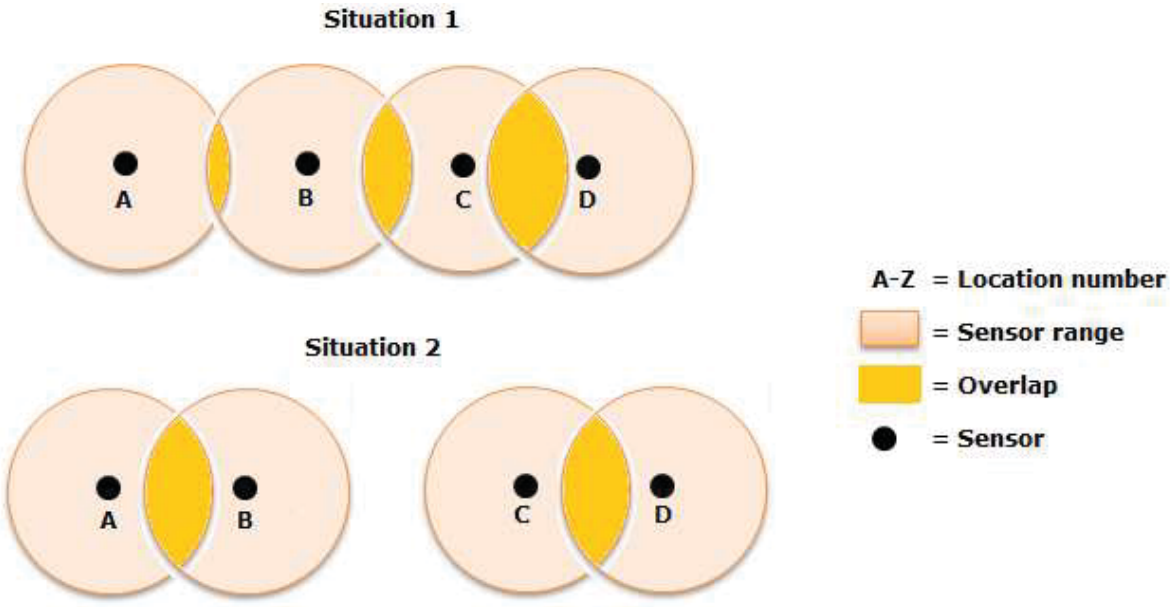


Figure 25: Correlation between multiple sensors and correlation between two sensors

In order to obtain situation 2 it is checked whether the first location in (4-13) has previously been allocated to be the best match according to Equation (4-15). If this is the case, no attempt to find a better correlated sensor is made. If the correlation between these locations is considered insufficient, another better correlated sensor can be found. This can be determined by assessing if the correlation is within a certain range. In Figure 24 a correlation of 0.7 is chosen as a lower boundary for this range.

As a summary the algorithm for de-correlation is redefined taken into account the proposed modifications as mentioned above.

1. The locations are ordered in descending order with respect to the information content
2. Weights according to Equation (4-15) are calculated for every location except the first location, If the first location has not been used for correlation with another location or this correlation is not within the accepted range. The location with the largest weight, has this weight assigned
3. All except this location and the first location have their weight assigned according to Equation (4-14).
4. The first sensor is taken out of the set, and step 1 is repeated until there are no more sensors left.

When the total number of sensors to be placed is increased, the information content of the last sensors will become less relevant to the total score of a set of locations. This is due to the fact that the de-correlation procedure has been performed multiple times and the weight for the last sensors will already approach zero.

4.2.4 Sequence of operations

To summarise section 4.2.3, the different components introduced have been put in order in Figure 26. This scheme assumes that the measuring accuracy and frequency are fixed. If the rank of the Jacobian matrix is (nearly) rank deficit, indicating a low information content, information can be added by increasing the number of observations. This can be achieved by adding monitoring locations (increasing spatial density) or increasing the sampling frequency (increasing time density). Instead this sequence decreases the number of parameters to fit the information available from a fixed number of sensors placed on locations that provide the highest information content.

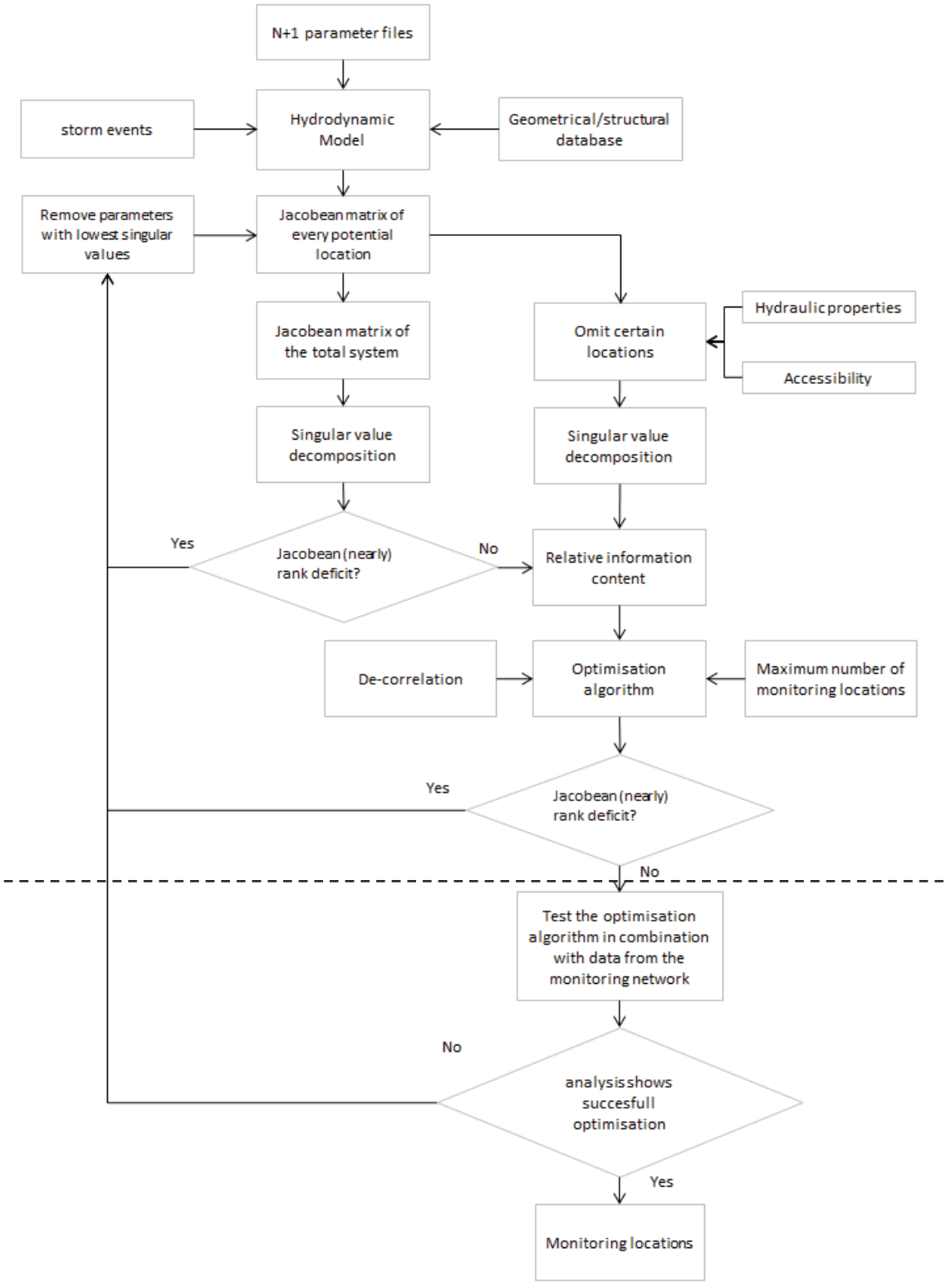


Figure 26: Working sequence for the design of a monitoring network

The part of the diagram which is below the dotted line is regarded as a test for the monitoring network. This test is used to judge the overall identifiability and sensitivity of the parameter set.

As has been mentioned previously, a singular value decomposition of the Jacobian matrix of the system reveals parameters eligible for optimisation. Parameters corresponding to (nearly) zero singular values are removed. Locations that maximise the information content on the remaining parameters are then incorporated in the monitoring network.

Chapter Summary

- A monitoring network is designed to meet a demand for information. This demand follows from the need for knowledge of a system, in order to make decisions concerning the system.
- A singular value decomposition of the Jacobian matrix is used to quantify the effect of a change in a certain parameter on the model results. This methodology is used to identify relevant parameters and to derive a set of monitoring locations.
- The objective of the de-correlation algorithm in the design process is to balance the overlap in the information collected, so that the combined information content of the locations is maximised while still retaining some overlap for the cross validation of data and sensor security.

5 Data assimilation

This chapter provides a more in depth view on the theory concerning data assimilation. The data assimilation method used in this thesis is introduced, and the performance of this method for a simple example is discussed.

The concept of data assimilation aims at combining uncertain models with uncertain measuring data in order to make the best estimate of the system state at a particular time at which observations of the systems are available (Hutton, et al., 2010). Sequential data assimilation is a specific type of data assimilation referring to the process where a new state prediction is produced when new observations are available to limit the deviation of the model with respect to the measured data. An example of variation in the state over time due to the availability of new measurements is presented in Figure 27.

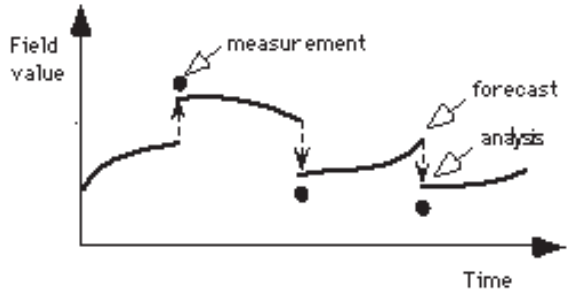


Figure 27: The application of sequential data assimilation on the state with respect to observations (Eskes, et al., 1998)

Although different data assimilation methods are available, this thesis focusses on the application of the Kalman filter for this purpose.

5.1 The Kalman filter

The classical Kalman filter (KF) is a popular algorithm for state estimation of linear systems. This technique was proposed by (Kalman, 1960), and is based on minimizing the variance of the estimation error (Simon, 2001). Noise affecting the system has to be uncorrelated in time, and have an average value of zero. The state is a description of the varying quantities of a system at a given time. In case of a simple linear system, matrix $\underline{\underline{A}}$ can be denoted as the model changing the state over time. evolution of the state over time is described by Equation (5-1).

$$\underline{x}_{k+1} = \underline{\underline{A}}\underline{x}_k + \underline{w}_k \qquad (5\text{-}1)$$

Where:

$\underline{x}_k$ = state at time k

$\underline{\underline{A}}$ = matrix containing information on the quantities affecting the state

$\underline{w}_k$ = process noise

The observations are related to the state by matrix *C* as seen in Equation (5-2) and can be represented by different quantities, for instance measured flowrates or water levels.

$$\underline{y}_k = \underline{\underline{C}}\underline{x}_k + \underline{z}_k \qquad (5\text{-}2)$$

Where:

$\underline{y}_k$ = observations at time k

$\underline{x}_k$ = state at time k

$\underline{\underline{C}}$ = matrix relating the state to the observations

$\underline{z}_k$ = measurement error

In the form of Equation (5-1) the state is not influenced by the observations, and the estimated observations are likely to increasingly differ from the real observations over time depending on how well the model describes the relevant processes. The Kalman filter makes a prediction of the state, and then compares the prediction of the observation with the real observation. The predicted state is corrected based on the error covariance matrices and the difference in the estimated observations and the measured observations. All the steps comprising the KF are presented in Figure 28. These steps can be categorized in forecast and analysis steps.

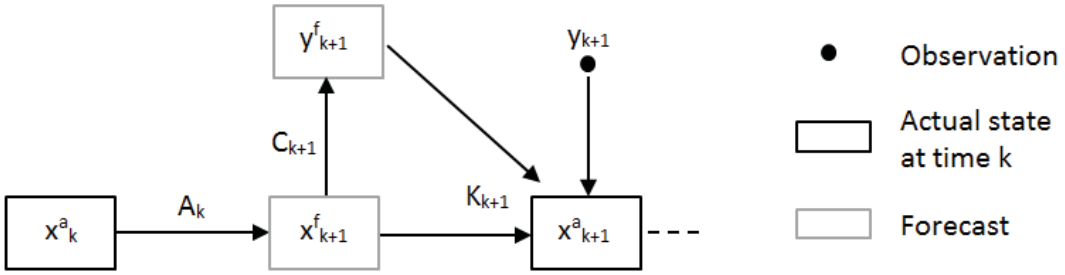


Figure 28: Kalman filter steps; the time step is denoted by k and the forecast by f.

It should be noted that the true state is never accessible, but information about the state is obtained through measurements. In the forecast step, an estimate of the state of the system and the observations for the next time step is made.

$$\underline{x}_k^a = \underline{x}_k^f + \underline{\underline{K}}_k \left(\underline{y}_k - \underline{y}_k^f \right) \qquad (5\text{-}3)$$

Where:

$\underline{x}_k^a$ = analysis estimate of the state

$\underline{x}_k^f$ = forecasted state

$\underline{\underline{K}}_k$ = Kalman gain matrix

$\underline{y}_k$ = observation

$\underline{y}_k^f$ = prediction of the quantity to be observed

Equation (5-3) is the main component of the analysis step and consists of two terms. The first term is the forecasted state, which is adjusted by the second term. If the difference between the forecasted and measured observations increases, the influence of the forecasted state decreases. This is almost trivial, since a large deviation indicates that the forecasted state does not approach the actual state. The Kalman gain is computed at each time step and depends on the model state error covariance and the measurement error covariance as stated in (5-4).

$$\underline{\underline{K}}_k = \underline{\underline{P}}_k \underline{\underline{C}}^T \left(\underline{\underline{C}} \underline{\underline{P}}_k \underline{\underline{C}}^T + \underline{\underline{S}}_z \right)^{-1} \qquad (5\text{-}4)$$

Where:

$\underline{\underline{K}}_k$ = Kalman gain matrix

$\underline{\underline{P}}_k$ = model state error covariance matrix

$\underline{\underline{C}}$ = matrix relating the state to the observation

$\underline{\underline{S}}_z$ = measurement error covariance matrix

From Equation (5-4) a number of characteristics of the Kalman filter can be deduced; when the measurement noise is increased the Kalman gain will decrease, therefore giving less weight to the observations, and making the state more dependent on the forecasted state. If the model state error covariance is large the influence of measurement errors is limited, thereby increasing the effect new observations have on the state. The model state error covariance is updated by:

$$\underline{\underline{P}}_{k+1} = \underline{\underline{A}}\underline{\underline{P}}_k \underline{\underline{A}}^T + \underline{\underline{S}}_w - \underline{\underline{A}}\underline{\underline{P}}_k \underline{\underline{C}}^T \underline{\underline{S}}_z^{-1} \underline{\underline{C}}\underline{\underline{P}}_k \underline{\underline{A}}^T \qquad (5\text{-}5)$$

Where:

$\underline{\underline{P}}_k$ = model state error covariance matrix

$\underline{\underline{A}}$ = matrix containing information on the quantities affecting the state

$\underline{\underline{C}}$ = matrix relating the state to the observation

$\underline{\underline{S}}_w$ = process noise covariance matrix

$\underline{\underline{S}}_z$ = measurement error covariance matrix

An example of the Kalman filter can be found in Annexe V.

5.2 The ensemble Kalman filter

Although the KF performs well for linear problems, it is found that most hydrological problems are non-linear in nature therefore requiring more advanced techniques (Drécourt, 2004). The Extended Kalman Filter (EKF) was introduced (Kalman & Bucy, 1961), and solves this problem partly by local linearization of non-linear systems. For systems with strong non-linear dynamics the EKF is found to be unsuccessful (Burgers, et al., 1998).

The Ensemble Kalman Filter (EnKF) is introduced by (Evensen, 1994) and integrates an ensemble of states forward in time. The spread of the model ensemble is used to represents the model error covariances, thereby eliminating the necessity to calculate the Jacobian. In Figure 29 the steps of the EnKF are shown for an ensemble size of 3. The linear system has been replaced by the model *f*, where *x* is the state of the system and *u* is the known input (e.g. rainfall measurements, geometry and known parameters).

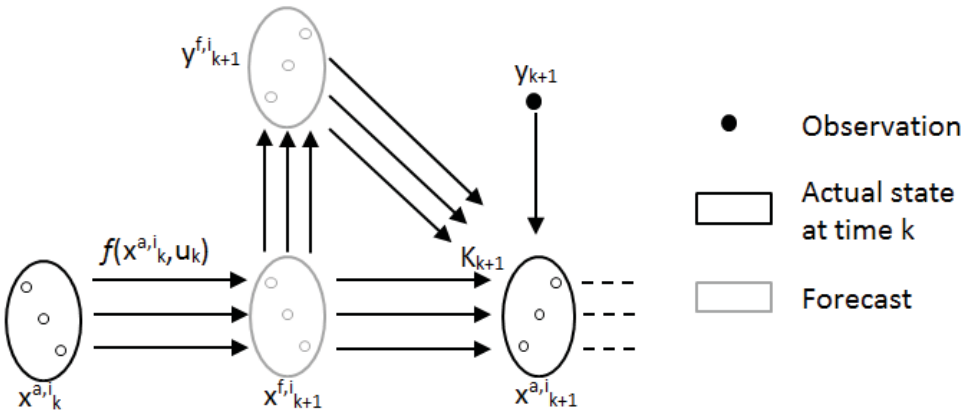


Figure 29: EnKF steps for an ensemble size i = 3

For each ensemble member the actual and forecasted state is calculated, and the forecast ensemble mean is regarded as the best estimate of the state. The Kalman gain and the measured observations are identical for each ensemble member. It is assumed that all probability distributions involved are Gaussian (Mandel, 2006). Since the probability distribution of the actual state is based on an ensemble of state estimates, and increase in the ensemble size will result in a better estimation of the true state. (Gillijns, et al., 2006) performed model runs with different ensemble sizes for various examples and used the mean squared error to assess the performance of the EnKF. It was found that an ensemble size of 50 to 100 is sufficient for most problems. A minimum ensemble size of 100 is reported by (Leeuwenburgh, 2005) for complex ocean and atmospheric models. The same article states that large ensemble sizes are too computationally expensive with respect to added benefits. Equations (5-6) and (5-7) describe the analysis step and forecast step respectively.

$$\underline{\underline{K}}_k = \underline{\underline{P}}^f_{xy,k} \left(\underline{\underline{P}}^f_{yy,k} \right)^{-1}$$

$$\underline{x}^{a,i}_k = \underline{x}^{f,i}_k + \underline{\underline{K}}_k \left(\underline{y}_k + \underline{v}^i_k - \underline{y}^{f,i}_k \right) \qquad (5\text{-}6)$$

$$\underline{\bar{x}}^{a}_k = \frac{1}{q} \sum_{i=1}^{q} \underline{x}^{a,i}_k$$

Where:

$\underline{\underline{K}}_k$ = Kalman gain matrix

$\underline{\underline{P}}^f_{xy,k}$ = forecast state error covariance

$\underline{\underline{P}}^f_{yy,k}$ = error covariance of the measurements

$\underline{x}^{a,i}_k$ = analysis estimate of the state for ensemble member *i*

$\underline{x}^{f,i}_k$ = forecasted state for ensemble member *i*

$\underline{v}^i_k$ = zero-mean random variable with a normal distribution

$\underline{y}_k$ = observation

$\underline{y}^{f,i}_k$ = prediction of the quantity to be observed

$\underline{\bar{x}}^{a}_k$ = state ensemble mean

q = ensemble size

i = ensemble number

$$\underline{x}_{k+1}^{f,i} = f\left(\underline{x}_k^{a,i}, \underline{u}_k\right) + \underline{w}_k^i$$

$$\underline{\bar{x}}_{k+1}^{f} = \frac{1}{q}\sum_{i=1}^{q} \underline{x}_{k+1}^{f,i}$$

$$\underline{\underline{E}}_k^f = \left[\underline{x}_{k+1}^{f,1} - \underline{\bar{x}}_{k+1}^{f} \cdots \underline{x}_{k+1}^{f,q} - \underline{\bar{x}}_{k+1}^{f}\right]$$

$$\underline{\underline{E}}_{y,k}^a = \left[\underline{y}_k^{f,1} - \underline{\bar{y}}_k^{f} \cdots \underline{y}_k^{f,q} - \underline{\bar{y}}_k^{f}\right] \qquad (5\text{-}7)$$

$$\underline{\underline{P}}_{xy,k}^f = \frac{1}{q-1}\underline{\underline{E}}_k^f \left(\underline{\underline{E}}_{y,k}^f\right)^T$$

$$\underline{\underline{P}}_{yy,k}^f = \frac{1}{q-1}\underline{\underline{E}}_{y,k}^f \left(\underline{\underline{E}}_{y,k}^f\right)^T$$

Where[2]:

$f\left(\underline{x}_k^{a,i}, \underline{u}_k\right)$ = non-linear system with known input u_k

$\underline{w}_k^i$ = stochastic forcing representing model errors

$\underline{\underline{E}}_k^f$ = ensemble error matrix

$\underline{\underline{E}}_{y,k}^a$ = ensemble of output error

For the complete derivation of the EnKF, the interested reader is referred to (Burgers, et al., 1998) or (Evensen, 2003).

5.2.1 Example of the EnKF for a simple reservoir model

In order to illustrate the application of the EnKF, a reservoir model is introduced. The model is schematically presented in Figure 30 and is governed by the mass balance in Equation (5-8). The inflow is precipitation on the surface of the reservoir, while the outflow is controlled by a weir.

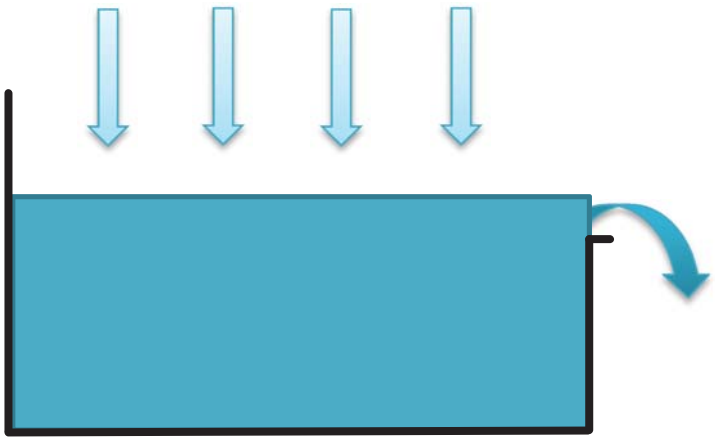


Figure 30: Reservoir model

[2] For the declaration of the other variables, see Equation (5-6)

$$\frac{dh}{dt} = I - \frac{B \cdot \alpha (h-\theta)^{\beta}}{A} \qquad (5\text{-}8)$$

Where:

h	= water level in reservoir	[m]
α	= weir coefficient	[$m^{(2-\beta)}/s$]
β	= weir power	[-]
A	= constant surface area of the reservoir	[m^2]
I	= rain depth	[m]
B	= width of the weir	[m]
θ	= crest level	[m]

The weir controlling the outflow will only function when the water level is above the crest level. The two parameters to be optimized are the weir coefficient α and the power β, and the water level in the reservoir is the observation. Since only parameters related to the weir are optimized, any water level below the crest level will not yield any information useful for data assimilation.

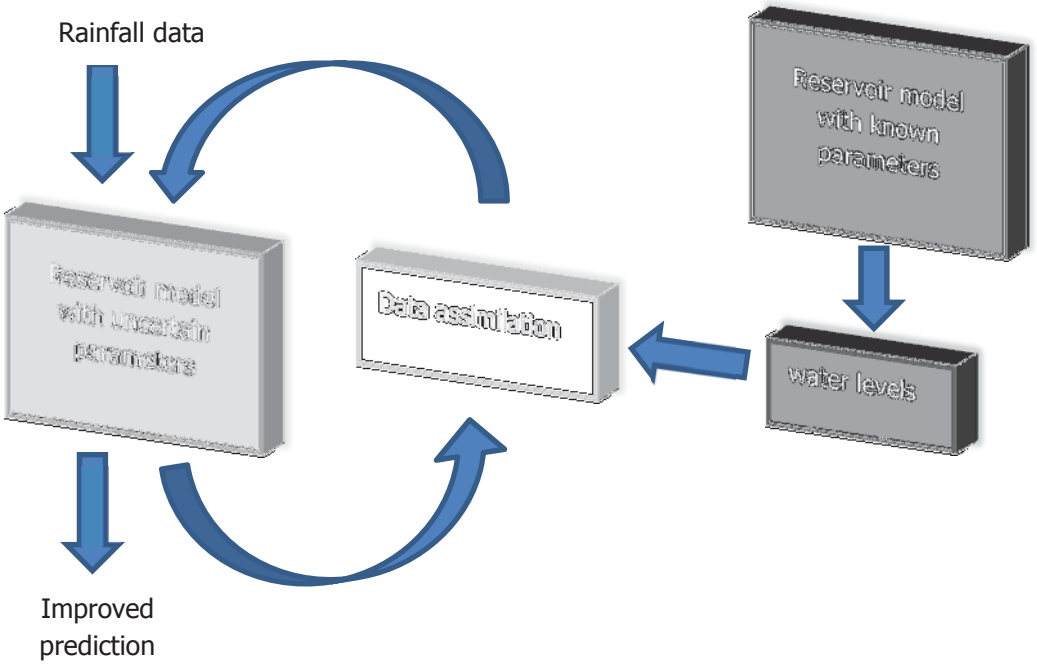


Figure 31: Principle of the twin model data assimilation concept

In order to test the performance of the EnKF a twin model experiment is conducted, as demonstrated in Figure 31. This comprises a model with known parameter values α and β to produce the observations that are assimilated in to a copy of the model in order to estimate the parameter values. This setup makes it possible to compare the parameter values derived by applying data assimilation to the values used to create the observations. It should be noted that successful assimilation for a twin model experiment does not guarantee the method to have the same performance for real measured observations, since that requires the assimilation method to compensate for processes not incorporated in the model as well.

For this example the time independent values for α and β used to create the observations are 0.4 and 1.4 respectively, and the crest level is 1.3 meter. To test the robustness of the EnKF, the initial parameter set of the assimilated model is incorrect (α = 1 and β = 2.5). The model is run for 60 time steps with a sampling interval of 2 seconds and an ensemble size 100. Precipitation is artificially

generated from an uniform distribution and the initial water level is below the crest of the weir. The outcome of the model is presented In Figure 32; the top graph illustrates the evolution of the parameter values over time, while the bottom graph shows the difference between the measured water level and the predicted water level by the EnKF. During the first time steps the parameter values do not vary, corresponding to the water level being below crest level. After this period, the deviation between the modelled water level and the measured water level is the largest. As the parameter values converges to the true values, the deviation between the measured water level and predicted water level rapidly decreases. After 20 time steps, water levels are simulated with error margins beneath 5 centimetres.

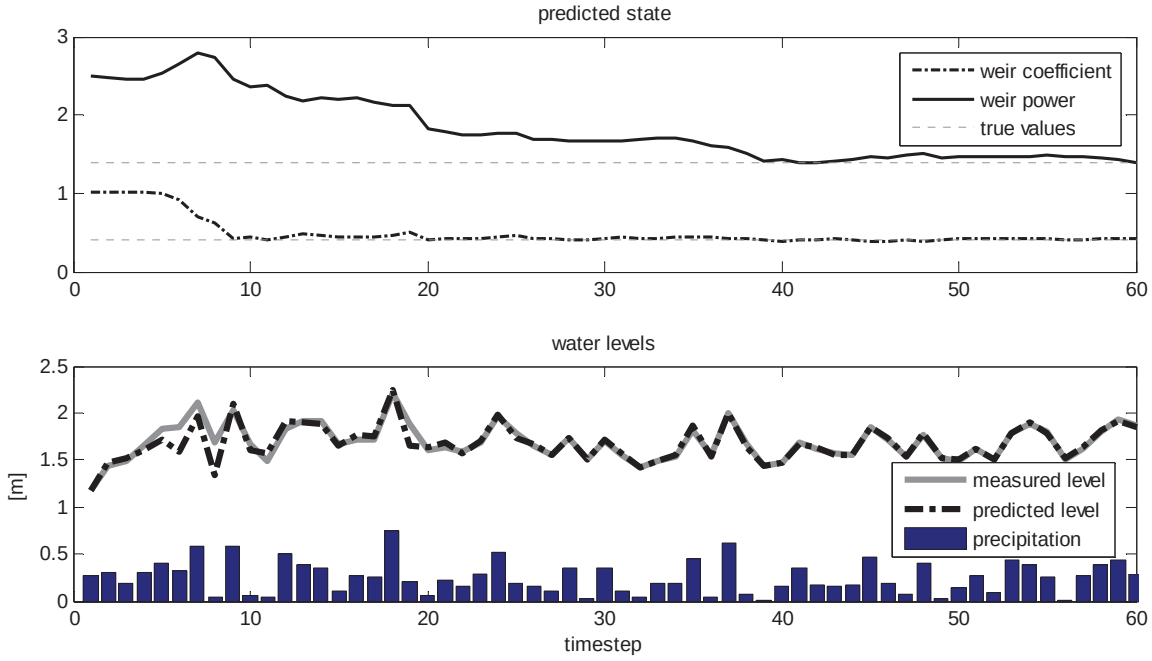


Figure 32: application of the EnKF on the reservoir model

There is a difference in the time the algorithm needs to reach the true parameter values, and the deviation from the true values over time after this moment. The weir power needs more time to approximate the true value, indicating the system is less sensitive to this parameter. This is also reflected in the water levels that hardly deviate from the measured water levels in the timeframe when the weir coefficient is stable and weir power is not.

A simulation with constant precipitation is also carried out, where the parameters values found diverged from the true values (α and β increased). This is logical, since more than one set of parameter values will yield the same water level when the intensity of the storm applied is time independent.

Chapter Summary

- The ensemble Kalman filter (EnKF) is the data assimilation method used in thesis. This method is known to perform well for non-linear problems.
- An ensemble of model states is integrated forward in time in order to estimate the model error covariances, therefore the number of ensemble members will influence the quality of the estimation of the true state.
- For a simple reservoir model in Matlab® the parameter values derived by the EnKF converge to the true parameter values, despite the incorrect initial estimation of these values.

6 Measuring setup case study

The methodology for the design of a monitoring network elaborated in chapter 4 is applied to the case study in this chapter. In order to reduce the computation time needed to find a set of monitoring locations, the application of a genetic algorithm is investigated. The set of parameters to be optimised is derived by analysing a singular value decomposition of the Jacobian matrix. Furthermore, the evolution of the information content with increasing sampling frequencies is analysed in order to derive a measuring frequency.

6.1 Omitted locations

A survey is conducted in order to identify locations not fit for placing and maintaining monitoring equipment. In Figure 33 four types of areas are presented, that are excluded for the placement of monitoring equipment. From the top left in a clock wise direction: a square with a market and other events limiting accessibility and increasing fouling, a high traffic intensity road introducing safety issues, a narrow alley and manholes (partially) on parking spaces. The latter is frequently present next to the canals where parking spaces are often located along the quayside.

Figure 33: Types of area identified to be less suitable for the placement of monitoring equipment (also see the above description)

In the surface water system, culverts are excluded from the list of possible measuring locations based on accessibility.

6.2 Genetic algorithm for de-correlation

Not taking into account the omitted locations, the surface water and sewer system has approximately 1100 potential monitoring locations. If the budget allows for the placement of ten sensors, this results in $7.05 * 10^{23}$ possible combinations according to Equation (6-1).

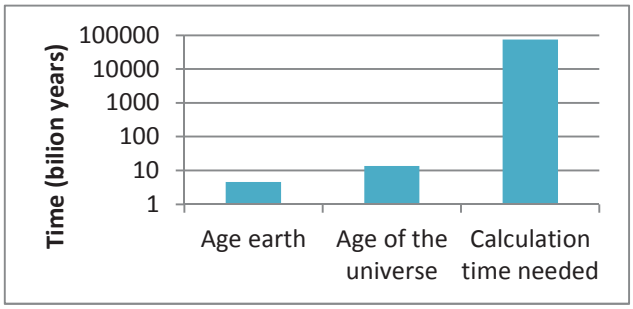
$$Combinations = \frac{n!}{(n-r)! \cdot r!} \qquad (6\text{-}1)$$

Where:

n = number of sensors to be placed

r = number of monitoring locations

Using a 8 GB RAM, i7-2630 QM processor circa 300 combinations can be processed per second, which can be translated to $7.45*10^{13}$ years needed to calculate the information content for all combinations using the de-correlation procedure as stated in chapter 4.



To get an idea of the spread in the scores assigned by the de-correlation procedure, 100 locations were taken from a uniform distribution. If 5 locations can be monitored this results in 75,287,520 possible combinations according to Equation (6-1). The empirical cumulative distribution function of the result is shown in Figure 34. The best set of locations has a total information content of 0.3742. As seen in Figure 34 the number of combinations that comes even close to this value is very limited, and clearly demonstrates the inefficiency of computing the total information content for all locations.

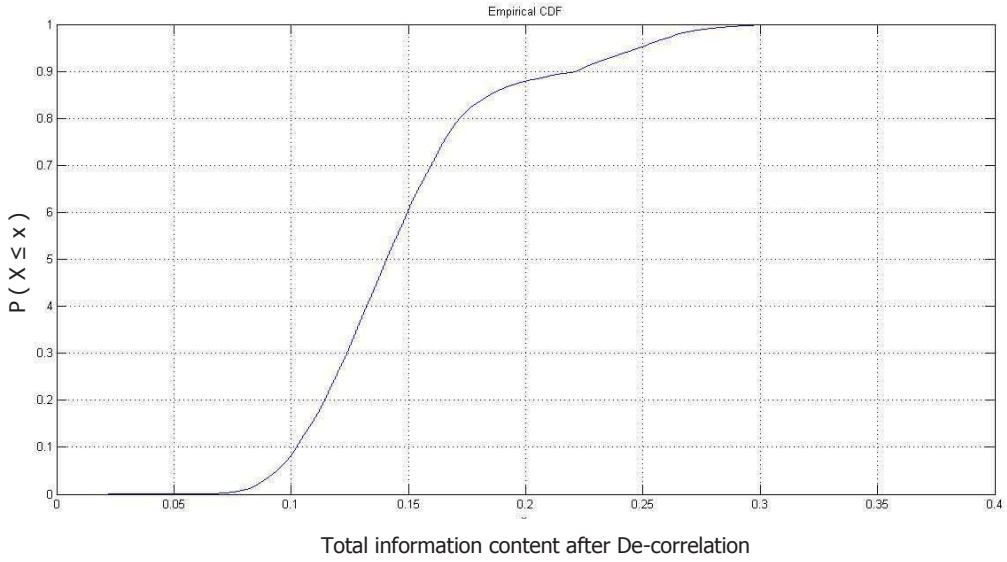


Figure 34: empirical cumulative distribution function for 100 locations

In order to reduce the computation load the application of a Genetic algorithm (GA) is investigated. The concept of GA was introduced by (Holland, 1975) to simulate the process of evolution. Individuals with better properties have a better survival chance and are therefore more likely to produce offspring's (survival of the fittest). Therefore, over several generations the individuals with these properties will become more dominant inside the population. This is referred to as natural selection.

It is evident that after a large number of generations the whole population consists of individuals with more or less the same properties. New traits are introduced in the population through random mutations. If these mutations result in individuals with better properties natural selection will spread these traits through the population over several generations, while mutations that cause less beneficial properties will gradually fade from the population. A simple example is shown in Figure 35,

where bacteria which are more resistant to antibiotics have a better survival chance and therefore come to dominate the population.

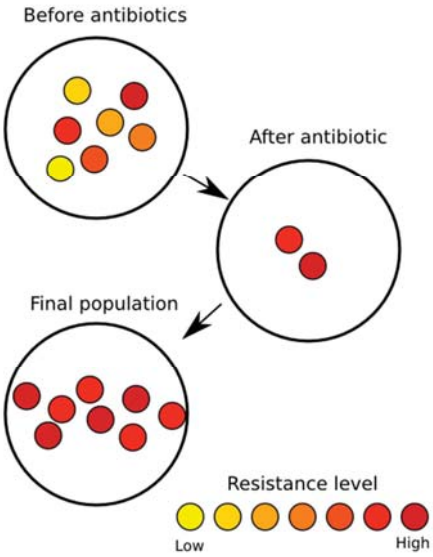


Figure 35: increasing antibiotic resistance due to natural selection (Urbano , 2010)

(De Jong, 1975) first started using a GA to solve an optimisation problem. Since then GA has been widely used in different fields including urban drainage, e.g. (Clemens, 2001), (Langeveld, 2004), (Rauch & Harremoës, 1999) and (Di Pierro, et al., 2005).

If the de-correlation procedure is referred to as the objective function which needs to be maximized, the maximum value or global maximum of this function is the set of locations which has the highest score. The GA has the ability to find different maxima in a large search space without getting stuck in a local maximum (see Figure 36), however since the process is stochastic the results may vary and the algorithm may have difficulties finding the global maximum.

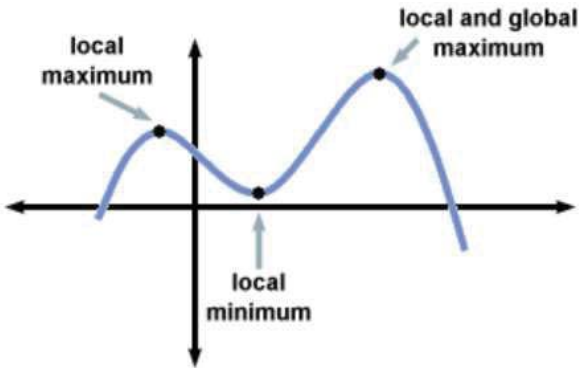


Figure 36: Shape of a objective function for a single parameter (Sparknotes, 2012)

An outline of the GA is given in the flowchart presented in Figure 37. The population consists of a predefined number of individuals. Each individual has a number of genes equal to the number of parameters that are required to be optimized, which in this case is the number of sensors to be placed. The fitness of each individual is calculated using the de-correlation procedure.

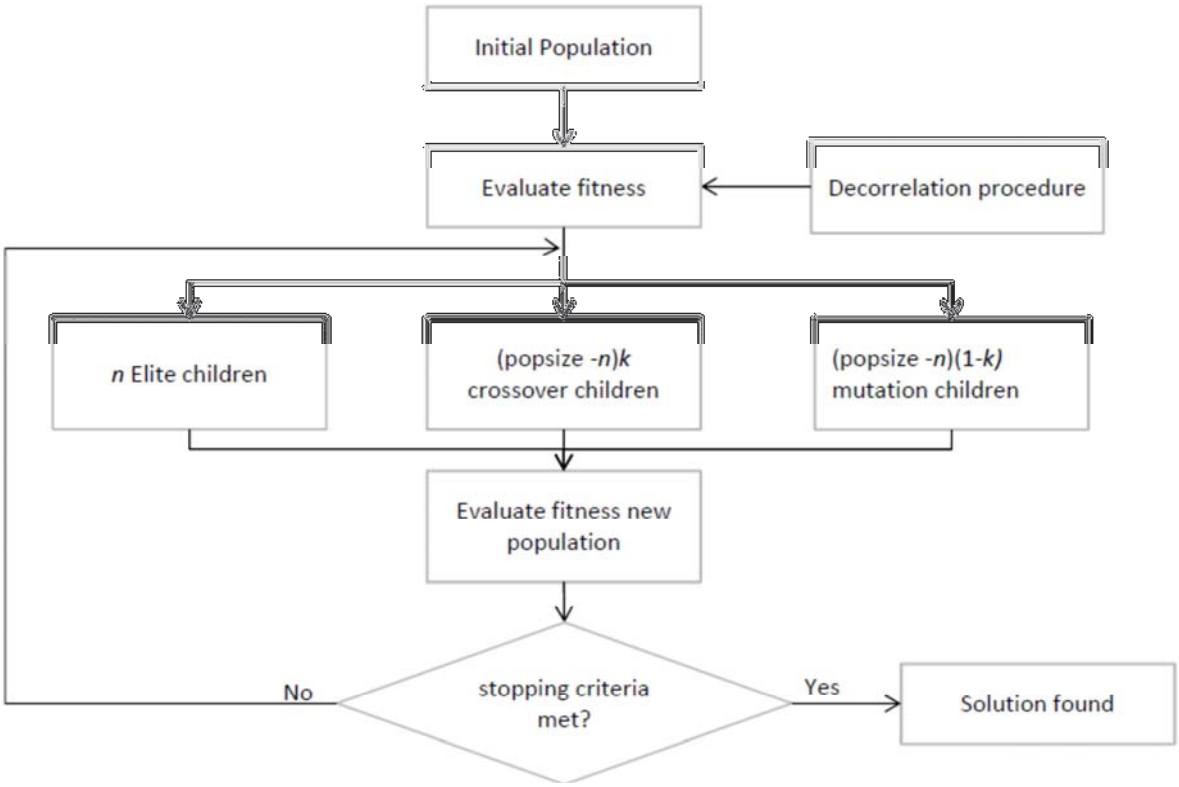


Figure 37: Genetic algorithm flowchart

The *n* individuals with the best fitness will be copied to the next generation, thereby preserving the best combinations of locations. From the remaining individuals a fraction (*crossover fraction*) is used to create crossover children (see Figure 38) by picking genes from two parents. The other children are created using the genes from a single parent where each gene has a certain chance to be mutated (*mutation rate*). In the end a new generation is created with the same size as the initial population.

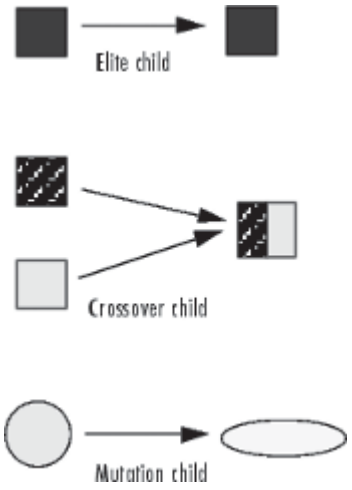


Figure 38: creation of the next generation by three kinds of children (Mathworks, 2007)

The optimization toolbox in Matlab® allows for the user to choose between different functions for creation of the population and children and the interested user is referred to the literature, e.g. (Haupt & Haupt, 2004), for a more in depth explanation. However a few remarks about the function of choice should be made taken into account the nature of the optimization problem:

- All genes should be integer values (since they refer to location numbers)

– The variables are independent

The former means that every member of the initial population and the mutated children produced by the chosen set of functions should be integer values between 1 and *n* where *n* is the total number of locations in the system. The latter refers to the fact that for example if an arbitrary location 3 has a high information content this gives no information about the information content of location 2 or 4, this can also be seen in Figure 39. Therefore choosing a mutation function that picks integers near locations that have a high information content will not improve the result, and might even worsen the result since it narrows the search space. This implies that the algorithm is more likely to be successful if the mutation function picks the locations from an uniform distribution, and that a larger population is needed for integer problems compared to non-integer problems in order to find the global maximum in a range of locations with local maximums that can be as wide as one integer.

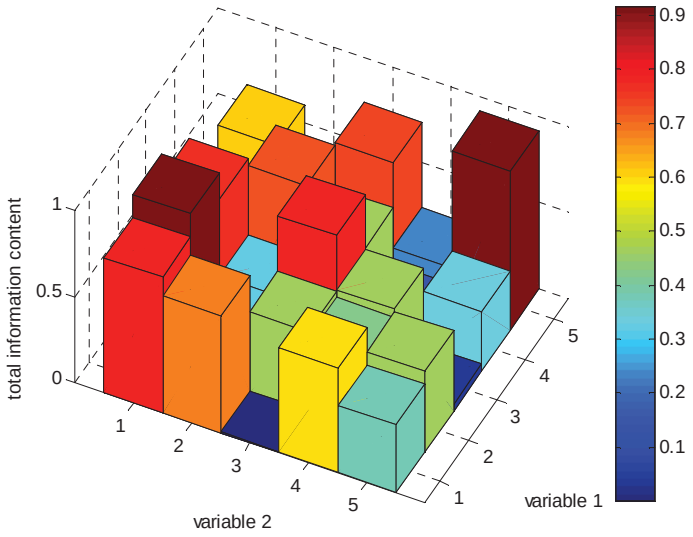


Figure 39: example of the total information content for two monitoring locations

Three important parameters to determine are the population size, the crossover fraction and the mutation rate. A good combination of these parameters will yield a solution that is (close to) the global maximum and is reproducible. Optimum values for these parameters depend on the fitness function (Mathworks, 2007) and therefore differ depending on the problem. Hereinafter two examples are elaborated with the same locations as was used to produce Figure 34. In the first example the standard mixed integer optimizer available in the Matlab optimisation toolbox is used. This optimizer uses a function for mutation and creation of the initial population that is bases on (Deep, et al., 2009). Test runs with different values for the population size and the crossover fraction have been made and repeated multiple times. The mean total information content is plotted in Figure 40 and the standard deviation in Figure 41. The algorithm is terminated when either 100 generations are computed, or the average change in the total information content is less than 10^{-20}.

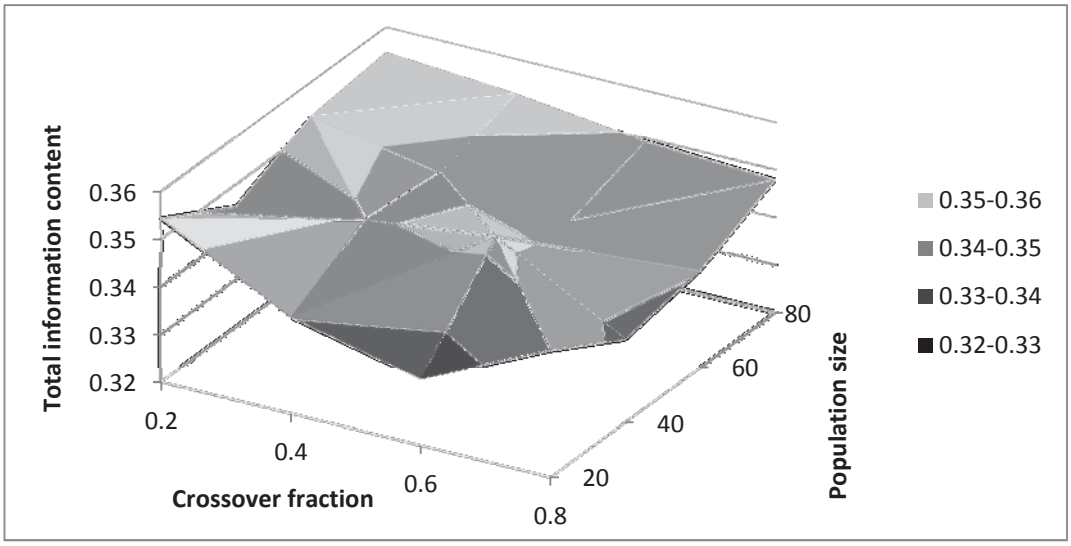


Figure 40: mean of the total information content using the standard mixed integer optimizer

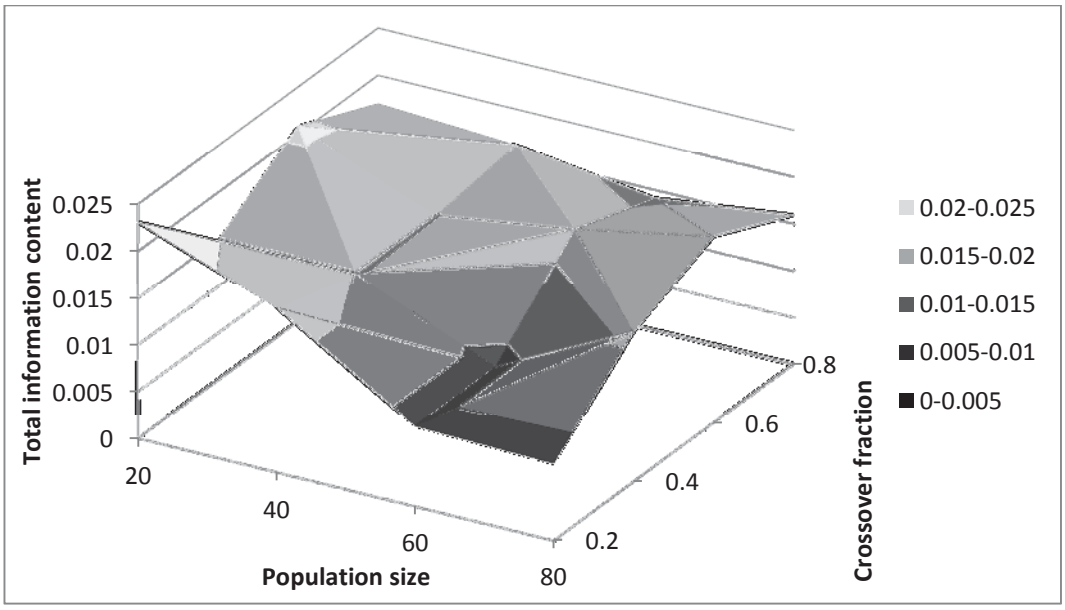


Figure 41: standard deviation of the total information content using the standard mixed integer optimizer

The plots together provide information on the quality of the result. The mean value should correspond to the maximum total information content found when computing all the possible combinations. A large standard deviation implies that the result found is not reproducible, which is influenced by the stochastic character of the algorithm. According to Figure 40 the best total information content found is for the largest population size and a crossover fraction of 0.2. It makes sense that the best solution is found when using the largest population, since a larger population means the solution space is searched more intensively. This is however at the cost of computation speed. The mean total information content found is not near the value found for the maximum total information content when computing all the possible combinations. The best mean total information content does however correspond to the lowest standard deviation according to Figure 41, indicating the result to be reproducible.

The second example utilizes a custom function for the mutation and creation of the initial population that has been created by the author. The creation function utilizes prior information given by the fact that the information content of each individual location is known. So if the number of locations in the

initial population is smaller than the total number of locations in the system, the creation function picks the locations with the largest information content for the initial population. However, if the number of locations in the initial population is larger than the total number of locations in the system, additional locations are picked from an uniform distribution.

The influence of the mutation function is determined by the mutation rate. The mutation rate is defined as the chance that a gene is mutated. The new value for the particular gene is also picked from an uniform distribution, and is set to be an integer value referring to a potential location. The results for different values of the mutation rate and the crossover factor is shown in the following plots. Each plot refers to a population size. The mean total information content is plotted in Figure 42 and the corresponding standard deviation in Figure 43.

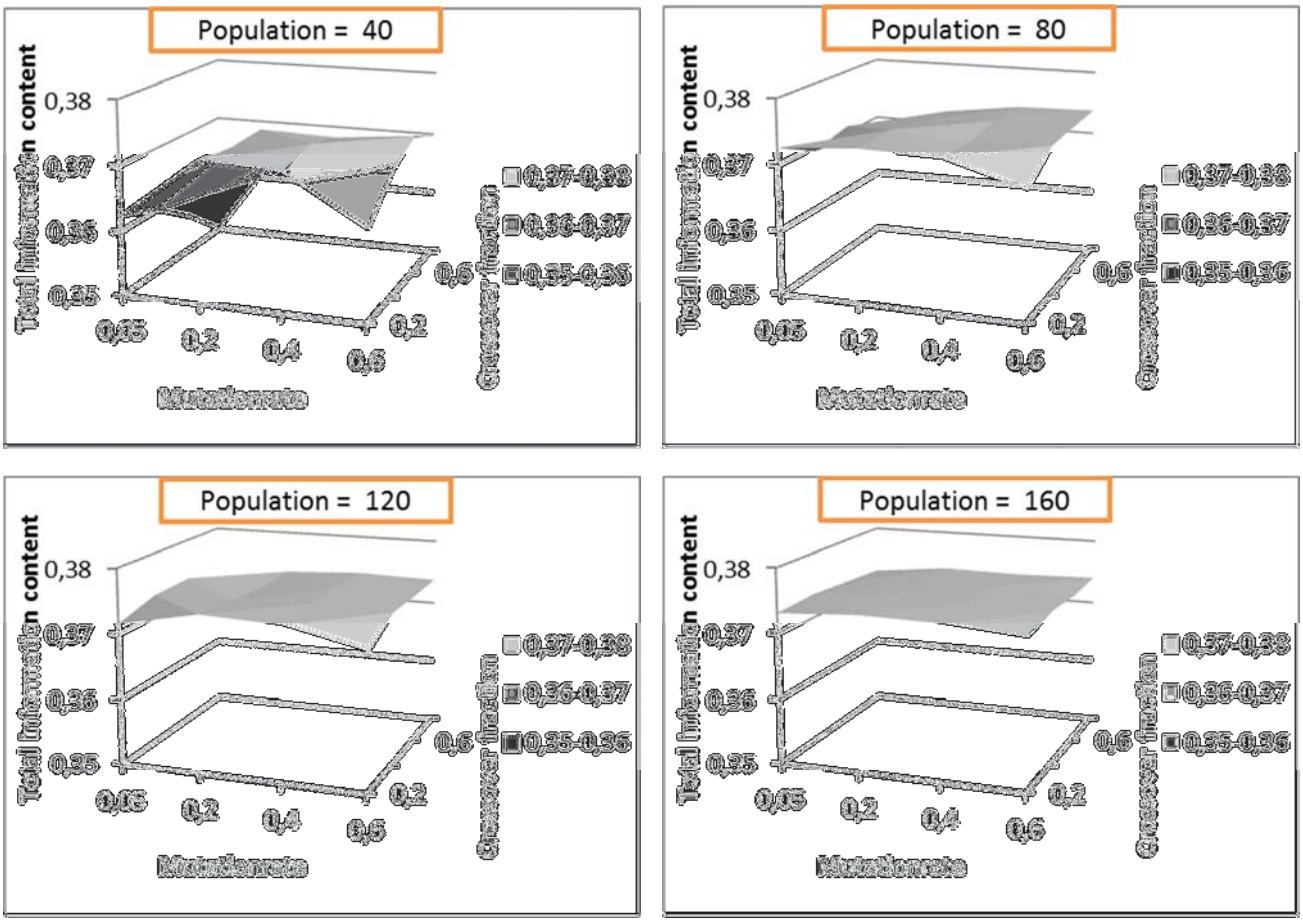


Figure 42: mean of the total information content using the custom optimizer

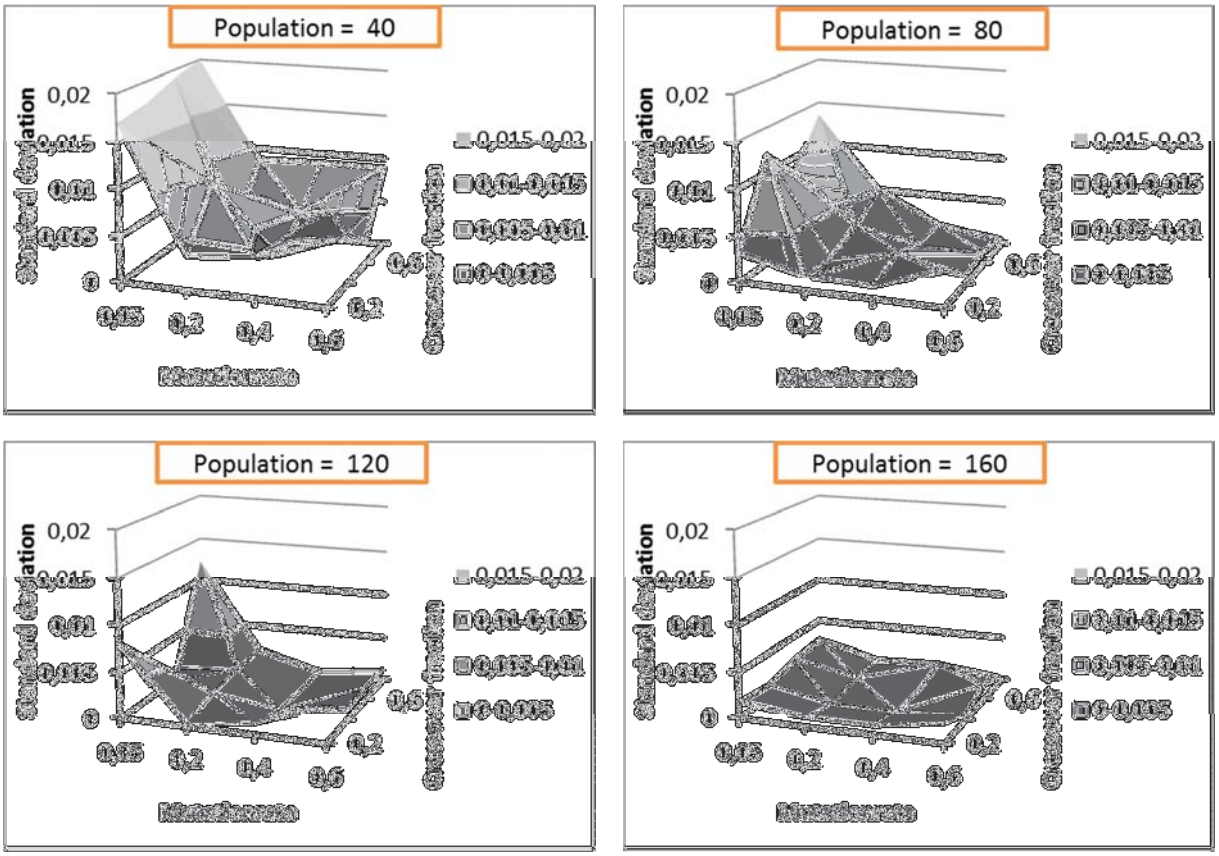


Figure 43: standard deviation of the total information content using the custom optimizer

The results in Figure 42 point out that although the computation load increases linearly with the population size the improvement of the result does not. This plot also points out that when a smaller population size is chosen, correct values for the mutation rate and the crossover fraction are vital for a good solution, while for larger population sizes the values chosen are less relevant. A similar result is observed for the standard deviation. The results obtained applying the custom optimizer function are closer to the maximum total information content for the same population size while having a smaller standard deviation compared to the results from standard mixed integer optimizer. Therefore one can conclude that the custom optimizer function is more efficient for this type of problem.

The next step is to apply the genetic algorithm to the case study. For the case study a monitoring network consisting of 31 sensors is designed. From these 31 sensors, 16 are already present on both sides of the adjustable weirs in the surface water system. Of the 15 remaining locations, three are fixed locations since these sensors provide information on the boundaries of the sewer system. This means that 12 monitoring locations need to be determined by the optimisation algorithm.

Using a mutation rate of 0.4 and a crossover fraction of 0.4 which are favourable according to Figure 42 and Figure 43 the population size is varied, since there are 1045 locations where a sensor can be installed in the case study area. Each population size is calculated 5 times in order to get an indication of the spread. Results from Figure 44 and Figure 45 indicate a population size of 1000 to be sufficient. Increasing the population size further will dramatically increase the computation time while not resulting in a much better set of locations.

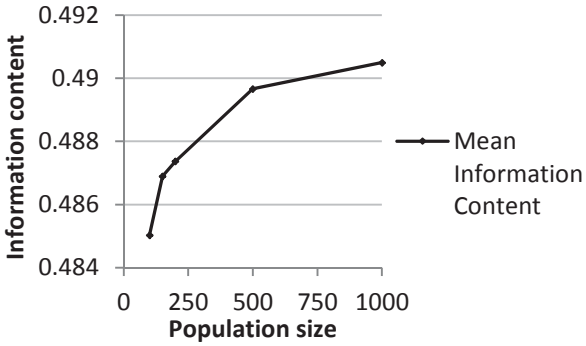


Figure 44: Mean information content of 31 monitoring locations chosen from 1045 locations where sensors can be installed

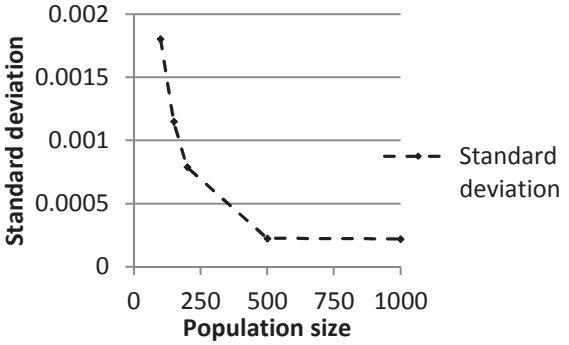


Figure 45: Standard deviation of 31 monitoring locations chosen from 1045 locations where sensors can be installed

6.3 Parameters for optimization

In this paragraph a set of parameters for the case study, eligible for optimisation is derived by using a singular value decomposition of the Jacobian matrix. According to (Ummels & Clemens, 1998)only process related parameters should be optimized, and that other parameters involved are considered to be fixed values. Since the majority of the parameters introduced in the model of the case study are weir coefficients, this group of parameters is first studied in more detail.

There are 49 weirs located in the sewer district of interest. This large number of weirs is inherent to the limited freeboard in the area, which cannot cope with high energy levels. As a first indication, the number of parameters optimized should be one or two orders of magnitude smaller than the number of measuring locations (Clemens, 2001). Therefore it is necessary to reduce the number of weir coefficients in order to keep the monitoring network cost effective. A reduction of weir coefficients can be obtained by omitting certain weir coefficients or by clustering weir coefficients together. The former method is suitable for weirs that have an insignificant effect on the system, for instance a weir that has a high crest level and therefore rarely overflows. The latter method can be applied for locations with similar flow patterns and geometry, or locations that cannot be identified separately.

Since the value of a weir coefficient does not only depend on the type of weir but also on the approach flow velocity, plan contraction and the crest rounding (Hager, 2010) one can conclude that not only the geometry of the CSO manholes need be similar to a large extend but also the feed and drainage structures of the CSO need to be similar. In this thesis, weir coefficients are not combined in this way since the model does not contain the necessary information required due to simplifications and conversions. Using the singular value decomposition discussed in chapter 4.2.3.3 weir coefficients that cannot be identified separately are clustered. Since two parameters cannot be identified separately when they influence the same area in time and space, it is expected that weirs in the same region with similar crest levels can be combined. A typical example in the Delft area can be found where inverted siphons are used to connect drainage areas separated by open water (see Figure 46). The CSO structures are located close together and have similar crest levels.

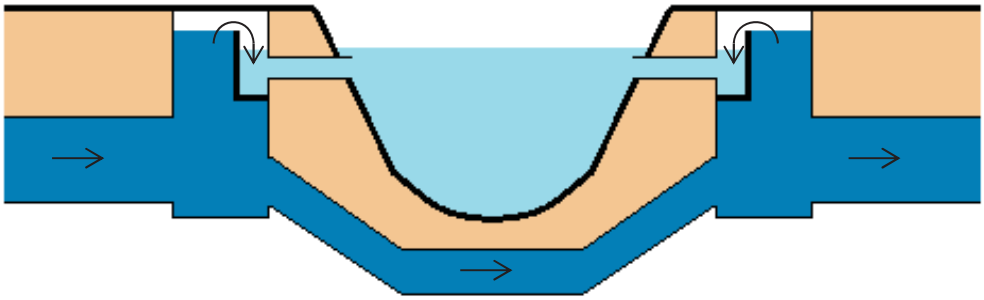


Figure 46: Typical inverted siphon construction under a city canal with a CSO on both sides

It was found that the amount of RAM available is limiting for the export of data from the hydraulic software package used. Therefore the number of parameters that can be studied for the amount of possible monitoring locations in the system is restricted to 28. Since the number of parameters related to the hydraulic behaviour of weirs is by far the largest, it is decided to first reduce this subset. For this purpose the Dutch design storm nr 2 is applied as hydraulic load. This storm event is characterized by a relative short duration, resulting in less data to process, with an intense peak of 50 l/s*ha at the end corresponding to a return period of 0.25 years. Since the model is not able to reproduce the processes related to pluvial flooding correct, a more intense storm event is not applied. The resulting singular values and eigen vectors of the network after the first round of clustering can be found in Annexe I. Several weirs are omitted from the set because they dominate eigen vectors corresponding to nearly zero singular values. A closer inspection shows that these weirs have relative high crest levels, or can be bypassed internally.

After several runs the parameter subset is reduced to a total of 14 weir coefficients, mainly due to the clustering of parameters. It is found that whether parameters can be identified separately, is more related to the crest levels than spatial distance, since some weirs located 750 meters apart cannot be identified separately. This means that the area of influence of a weir is large, resulting in overlap with another weir quickly. from this phenomenon one can deduce that the separate identifiability of weir coefficients is more related to measuring density in time than space.

Merging the weir coefficients with the friction parameters and runoff parameters yields the list of 27 parameters presented in Table 2.

Table 2: Set of parameters relevant to the applied model after clustering the weir coefficients based on identifiability

Runoff parameters	
(R1)	Linear reservoir constant closed flat
(R2)	Linear reservoir constant open flat
(R3)	Linear reservoir constant roof sloped
(R4)	Linear reservoir constant roof flat
(B1)	Storage closed flat
(B2)	Storage open flat
(B3)	Storage roof sloped
(B4)	Storage roof flat
(I1)	Infiltration open flat
(D1)	DWF (l/h*p)

Friction parameters	
(F1)	Pipe roughness
(F2)	Channel bed friction
(F3)	Friction bridges + culverts

Weir coefficients	
(W1)	Weir 3
(W2)	Weir 8
(W3)	Weir 21
(W4)	Weir 22
(W5)	Weir 23
(W6)	Weir 42
(W7)	Weir 49
(W8)	Weir 17+19
(W9)	Weir 4+11+9+10+18
(W10)	Weir 34+35+46
(W11)	Weir 30+32+37+29+33+36+38
(W12)	Weir 5+6+7+15+16
(W13)	Weir 13+14+20+31
(W14)	Weir 39+41+43+26+44+24+25

A singular value decomposition for this parameter set is calculated for three different storms. The number of storms used is limited by the available computer power, for the three applied storms the Jacobian has more than $121*10^6$ entries. The set of storms is chosen to be able to provide information on most of the parameters without pluvial flooding occurring. A description of the progress of the different storms over time can be found in Annexe II. Standard design storm 02 is chosen because of its high intensity peak, providing information on the most important weirs. The storm event of 24-07-11 has a higher total water depth, which is considered to be more critical for the surface water system and will give information concerning the importance of friction in the water system. The storm event of 19-01-12 has a lower peak intensity compared to standard design storm 02, but a longer duration and a higher total water depth.

The singular values and the corresponding eigenvectors of the network for these three storms are presented in Annexe III. Pipe roughness predominates the eigen vector corresponding to the largest singular value. Storage open flat and infiltration open flat both influence the same eigen vector, which can be explained by the fact that both parameters influence the total amount of water entering the system for that surface type. Remarkable is the interaction between W2 and several runoff parameters (see Figure 47). W2 is the weir partially blocking flow to the pumping station during storm events, one of the locations holding the most information. Because the crest level of this weir is the lowest in the system, it has a large influence on the system in a stage normally dominated by the runoff parameters.

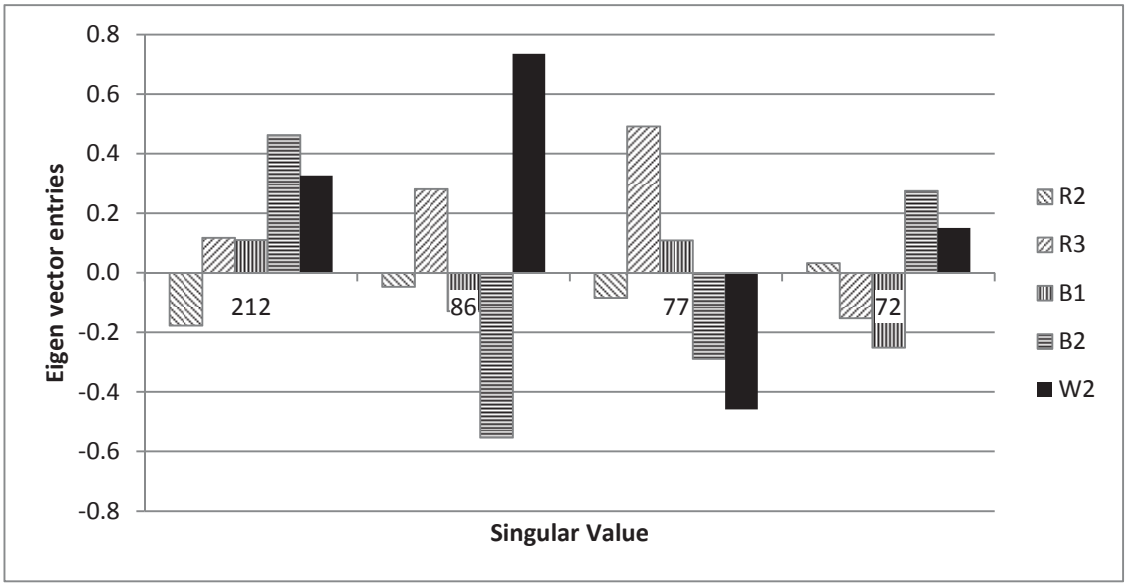


Figure 47: Values of the eigen vector for different singular values

After the genetic algorithm was applied for 15 new monitoring locations, a further reduction of the number of parameters was possible since information is now collected from the subset of locations, instead of the whole system. The resulting singular values and eigen vectors are presented in Table 3.

Table 3: Singular values and corresponding Eigen vectors for 15 locations selected by the genetic algorithm

	Singular values																		
	7.95E+03	2.73E+02	1.08E+02	8.66E+01	7.21E+01	4.24E+01	2.91E+01	2.66E+01	1.90E+01	1.71E+01	1.49E+01	1.46E+01	5.01E+00	3.48E+00	2.17E+00	1.59E+00	5.07E-01	2.14E-01	1.51E-01
Parameters	**Eigen vectors**																		
R1	0.0	-0.3	0.0	**-0.9**	0.1	-0.1	-0.1	-0.1	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
R2	0.0	-0.2	**0.9**	0.1	0.1	-0.2	-0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
R3	0.0	-0.1	0.2	-0.1	0.1	0.6	0.6	0.4	-0.1	0.4	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
R4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	-0.3	**0.9**	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
B1+B2	0.0	0.0	0.1	0.0	0.0	**0.7**	-0.4	**-0.6**	-0.2	0.1	0.1	-0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
B3	0.0	**-0.9**	-0.3	0.3	-0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
B4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	-0.1	0.0	0.0	**-1.0**	0.0	0.0	0.0	0.0	0.0	0.0
I1	0.0	0.0	0.0	0.0	0.0	0.2	**-0.7**	**0.6**	0.3	0.2	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
D1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**1.0**	0.0	0.0	0.0	0.0
W2	0.0	0.0	0.0	-0.1	0.0	0.3	0.0	0.4	-0.1	**-0.7**	-0.1	-0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0
W4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**1.0**	0.0
W7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**1.0**	0.0	0.0	0.0
W11	0.0	0.0	0.0	-0.1	0.0	0.0	0.1	-0.2	**0.8**	0.2	0.0	-0.5	0.1	0.0	0.0	0.0	0.0	0.0	0.0
W3+W5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**1.0**	0.0	0.0	0.0	0.0	0.0
W1+W6+W8+W10+W13	0.0	0.0	0.1	0.0	0.0	0.2	0.1	-0.2	0.4	-0.4	-0.4	**0.6**	0.1	0.0	0.0	0.0	0.0	0.0	0.0
W9+W12+W14	0.0	0.0	0.1	-0.2	**-1.0**	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
F1	**-1.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
F2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**1.0**	0.0	0.1
F3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	**-1.0**

In each column the largest entry of the eigenvector is in Bold, linking the largest singular value to a parameter. R3 does not dominate any of the singular values. Most of the parameters can be identified separately for the storms applied.

6.4 Monitoring locations

This section goes more into detail on the set of monitoring locations obtained by applying the genetic algorithm. Three sets of monitoring locations were obtained after running the genetic algorithm 12 times. Intuitively the set with the highest information content, corresponding to the largest sum of singular values was analysed. This set is characterised by a number of sensors close to CSO structures with the lowest crest levels. The eigen vectors determined by the dry weather flow and the channel bed friction are presented in Table 4.

Table 4: Selection of the singular values and corresponding eigen vectors determined by the DWF and channel friction

	Singular values	
	1.699863	1.430787
Parameter	**Eigen vectors**	
Dry weather flow sewer system	-0.7	-0.7
Channel bed friction	0.6	-0.7

The first column implicates that a decrease in the dry weather flow in the sewer system can be compensated by an increase in the channel bed friction in order to obtain the same model results, and means the parameters cannot be identified separately. This can be explained by looking at the characteristics of the system; due to the low crest levels an increase in the channel bed friction will force a larger volume of surface water in the sewer through the sewer overflow structures. The decrease in dry weather flow is compensated by this volume. However, the second column claims the opposite. This is due to the fact that the water levels in the channels are not increased during the whole simulation period, resulting in deviating model results in the remaining time steps. for the second best set of locations obtained by applying the genetic algorithm, these parameters can be identified separately. Therefore, this set of monitoring locations is used during the remainder of this thesis. The resulting singular values are already presented in Table 3, and the spatial distribution of the sensors is shown in Figure 48 and in more detail in Annexe IV.

Figure 48: Spatial distribution of the 12 sensors derived from the optimisation algorithm

By analysing the singular value decomposition of the individual monitoring locations, a distinction can be made in what information is collected from what location. It is found that location 1 in Figure 48 has the largest sum of singular values. This is due to the extra discharge from the nearby pressure main, which increases the effect that a change in certain parameters have on the water level. One could question the suitability of this monitoring location, since the pressure main can cause unwanted turbulence that can decrease the measuring accuracy. Runoff parameters R4 and R5 are more pronounced at other locations, which is in accordance with the distribution of the connected surface area. Surface water related friction is dominated by the monitoring location in the surface water, as would be expected. Information on weir related parameters is best obtained from monitoring locations near the corresponding weir. It could be considered to remove location 7; the total information content of this location is only 15% of the system average and it does not contribute to the overall system redundancy due to a round-off error as is explained in more detail in section 7.2.1. In general, the total information content of a location is not a valid criteria to judge whether a monitoring location can be removed from the system, since it is often found that these locations contribute significantly to the identifiability of a single parameter (e.g. weir coefficient). However for location 7 this is not the case.

6.4.1 Sensor correlation

Performance of the de-correlation algorithm is judged by analysing the correlation coefficients of the chosen monitoring locations. Two groups are distinguished; one group where a correlation>0.6 is promoted which consists of one value per sensor, and one group comprising the remaining values per sensor excluding the correlation with the location itself where a correlation>0.3 is punished. The results are presented in Figure 49.

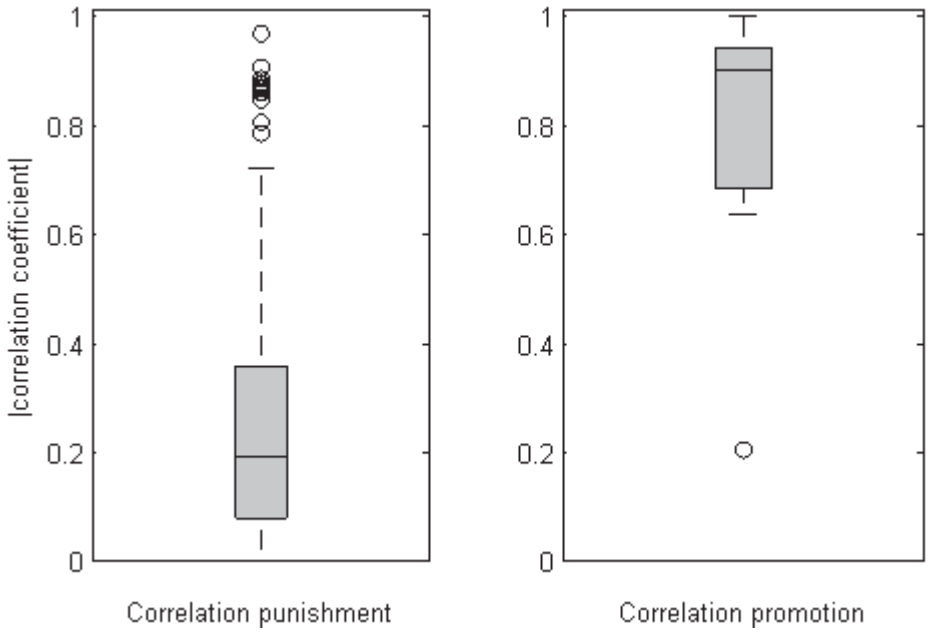


Figure 49: Box and Whisker plot of the correlations coefficients for the derived monitoring locations

The right plot contains the largest correlation value for each location, excluding the correlation each location has with itself. Therefore this plot can be used to judge the promotion of correlation with one other monitoring location for the purpose of data validation and redundancy. One can see that that the promoted correlations are all in the range indicated by (Harder, 2010) except for one outliner. This value is caused by a rounding off error in the correlation matrix that allows the sensor to be correlated with itself, causing the location with the highest correlation to be specified by the correlation-punishing function. this type of error can easily be addressed by implementing a function that rounds of the values before comparing it to the standard.

The remaining correlation values for the other locations are included in the left plot. This boxplot shows to what extend different sensors in the monitoring network collect the same information. it can be seen that approximately 75% of the values have a correlation smaller than 0.35 when a correlation threshold of 0.3 was applied, indicating diversity in the information collected. 4.5% of the values in this plot are outliners with a correlation larger than 0.71. Most of these outliners can be ascribed to locations with an information content > 99% of all locations, meaning that for these locations the information content outweighs correlation.

6.5 Measuring frequency

As mentioned in Chapter 2.3.1 Frequency domain analysis is applied to find an upper boundary for the measuring frequency. Dutch design storms served as hydraulic load, which are known for their high rain intensities resulting in a low characteristic timescale (Clemens, 2001). It was found that for most potential sensor locations no realistic upper bound for the sampling interval ($\Delta t \geq 1$ min) could be determined. This means that increasing the sampling frequency above once every minute can potentially result in an increased information content.

For the case study the singular values are determined for different sampling intervals. The singular values presented in Figure 50 are normalized with respect to $\Delta t = 1$ min. This graph provides insight in how the singular values decrease with an increasing sampling interval, and the variation in decrease over the different parameters values.

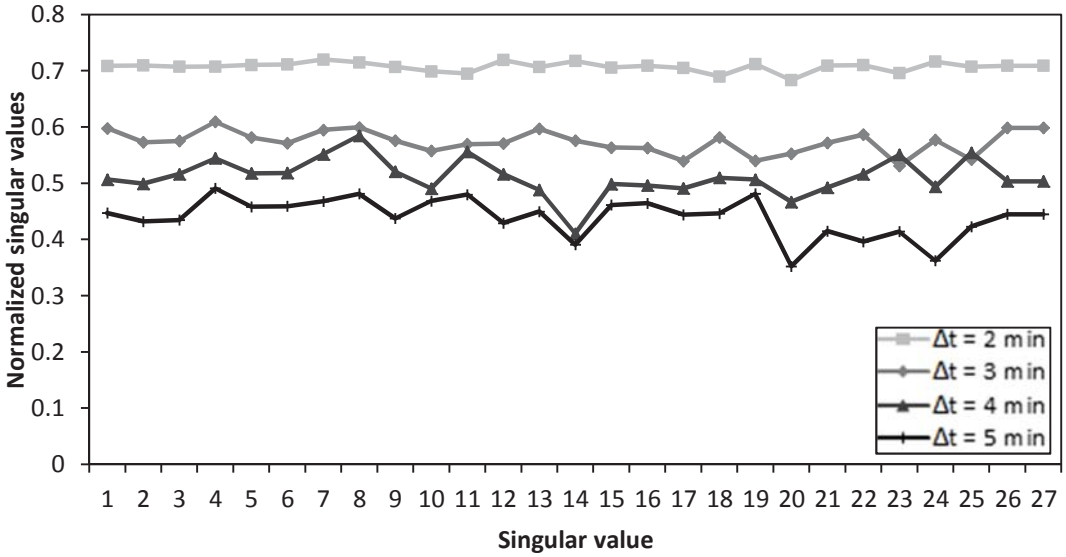


Figure 50: Singular values normalized with respect to Δt = 1 min for 27 different parameters

For larger sampling frequencies the decrease in singular values is equal over the whole set. When the sampling frequency is decreased, the variation in the normalized singular values increases. The normalized values for some weirs decrease faster with respect to the mean. This can be ascribed to the difference in characteristic time scales; during overflow mode the characteristic timescale is less, therefore making a small sampling interval necessary in order to obtain sufficient information to reconstruct the process. The weir coefficients that decrease less over increasing sampling intervals are found to have lower crest levels, therefore providing information over a prolonged period.

For two parameters, the sampling interval of four minutes has a larger information content than for a sampling interval of t minutes. These parameters refer to the surface water system, where the water levels have a small sinus like distortion. Due to the fact that four is no integer multiplication of three, the sampling interval of three minutes does not contain all the same points as the sampling interval of four minutes. Further examination of the Jacobian showed a period that had less peaks at the sampling interval of three minutes.

Even though Figure 50 shows the decrease of singular values for the whole system, it provides no insight in the distribution of this information over the different potential locations. If the decrease of the total information content is spread over only a limited amount of locations it is deemed acceptable, since only a small number of locations can be monitored.

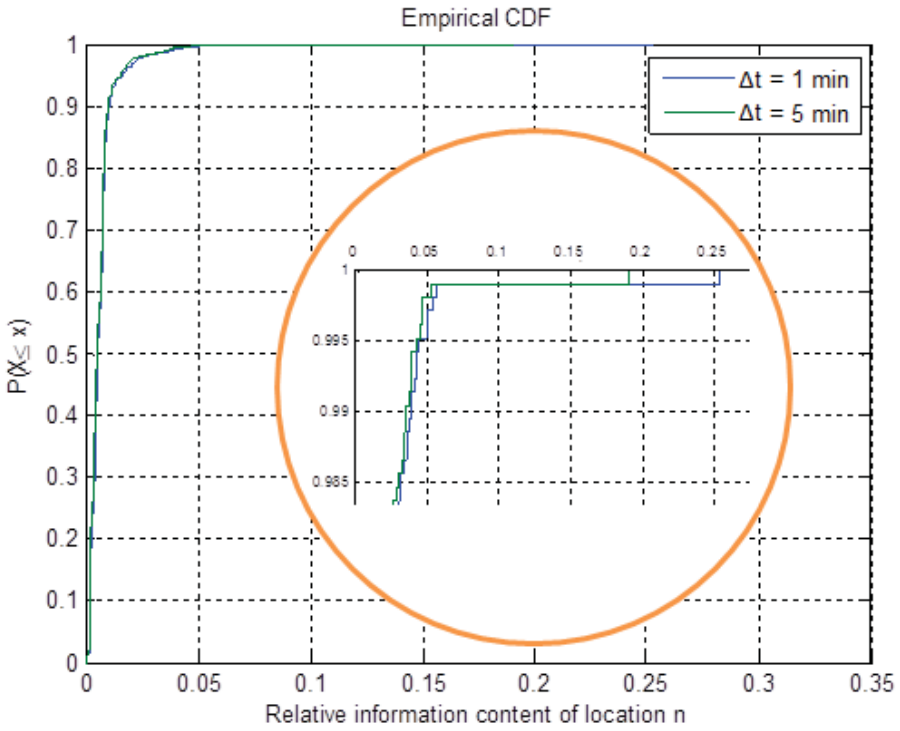


Figure 51: Cumulative distribution function of the relative information content, with enlargement of the top part

The data presented in Figure 51 shows that the distribution of the information for two different sampling intervals corresponds to a large extend, indicating an equal spread of the information content. An enlargement of the top 98.5% (corresponding to the best 15 locations) illustrates that for a smaller sampling interval the best locations hold a slightly higher portion of the total information content, especially for the best location.

Again the eigen vectors are studied in order to see if the different parameters can be identified separately when the sampling interval is increased. Standard design storm 02 is used, in order to minimize the characteristic timescales of the different processes. The lower boundary for the sampling interval was set to Δt = 1 min and increased with a time step of one minute. Between a sampling interval of one and three minutes no significant changes in the eigen vectors are found. It should be noted that the eigen vectors cannot be directly compared over different time steps, since the dominating parameter can change due to the fact that singular values for some parameters decrease faster with an increasing time step. Between the sampling interval of three and four minutes, the entries of the eigen vectors corresponding to the weir coefficients change substantially.

Table 5: Eigen vectors dominated by weir coefficients in descending order with respect to the singular values, Δt = 3 min

Parameter	Corresponding eigen vectors for Δt = 3 min						
W2	**-1.0**	0.0	0.0	0.0	0.0	0.0	0.0
W4	0.0	0.0	0.0	0.0	0.0	0.0	**1.0**
W7	0.0	0.0	-0.1	-0.2	0.0	**1.0**	0.0
W11	0.0	0.0	-0.3	0.1	**0.9**	0.0	0.0
W3+W5	0.0	0.2	-0.3	**0.9**	-0.2	0.2	0.0
W1+W6+W8+W10+W13	0.0	0.2	**-0.9**	-0.3	-0.2	-0.2	0.0
W9+W12+W14	0.0	**0.9**	0.2	-0.1	0.1	0.0	0.0

Table 6: Eigen vectors dominated by weir coefficients in descending order with respect to the singular values, Δt = 4 min

Parameter	Corresponding eigen vectors for Δt = 4 min						
W2	**1.0**	0.0	0.0	0.0	0.0	0.0	0.0
W4	0.0	0.0	0.0	0.0	0.0	0.0	**-1.0**
W7	0.0	0.0	0.1	0.0	-0.2	**-1.0**	0.0
W11	0.0	-0.1	0.2	-0.2	**0.9**	-0.2	0.0
W3+W5	0.0	0.3	**0.7**	**0.6**	-0.1	0.0	0.0
W1+W6+W8+W10+W13	0.0	**0.4**	**0.5**	**-0.7**	-0.2	0.1	0.0
W9+W12+W14	0.0	**0.8**	**-0.5**	0.1	0.2	0.0	0.0

Table 5 and Table 6 give an overview of the eigen vectors dominated by weir coefficients for the sampling interval of three minutes and four minutes respectively. The parameters in Table 5 can all to a large extend, be identified separately. This is not the case in Table 6, where the goal function is sensitive to an increase and/or decrease of the last three parameters.

It can be concluded that based on the decrease of singular values with an increasing sampling interval and the distribution of the information content over different locations no indication of the lower boundary for the sampling frequency can be obtained. However, analysis of the eigen vectors indicate that for this particular system the separate identifiability of the weir coefficients reduces significantly for a sampling interval higher than 3 minutes.

Chapter Summary

- Application of a genetic algorithm for the design of a monitoring network drastically reduces the computation time needed, while still approximating the maximum information content
- For the case study, the size of the parameter set is reduced by clustering parameters that cannot be identified separately. Due to the characteristics of the system, especially the number of weir related parameters can be reduced in this way.
- It is found that an increase in the sampling interval does not only result in smaller singular values, but has a significant impact on whether parameters can be identified separately. The latter is used to derive a lower boundary for the measuring frequency.
- The de-correlation algorithm is applied to the case study, and is successful in implementing the desired overlap in the monitoring network.

7 Data assimilation in urban drainage modelling

The EnKF introduced in chapter 5 is implemented in the hydrodynamic software package Sobek by the open interface standard OpenDA. OpenDA is supported by the TU Delft, Deltares and VORtech and comprises different data assimilation and calibration methods. Since the source code of Sobek is not altered, communication between both programs is achieved by reading/writing the input and output files of Sobek. This is called the blackbox model approach, since OpenDA controls the input and reads the output but has no further interaction with the model. In order for OpenDA to be able to read and write the necessary Sobek files, the drafting of several java files is required. The java files essentially govern the communication between both programs and are specific for the model used. OpenDA settings are specified in XML standard (Extensible Markup Language). Observations are generated using the twin model concept introduced in Chapter 5.2.1. Figure 52 shows the place of OpenDA and Sobek in this concept.

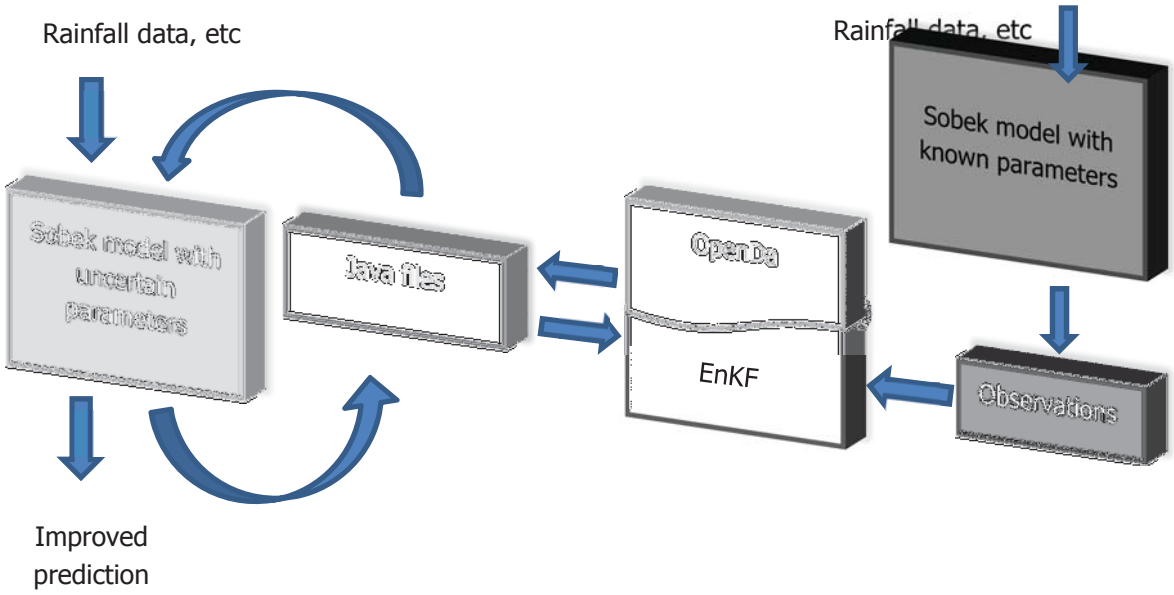


Figure 52: Twin model concept for Sobek and OpenDA

The runtime of the Sobek model needs to be controlled by OpenDA, because following every new observation a new Sobek run is started in order to implement the updated state. This principle is schematized in Figure 53. The top plot shows a normal simulation, and the bottom plot shows the same simulation period divided over five individual simulations. These simulations are denoted as sub runs in this thesis.

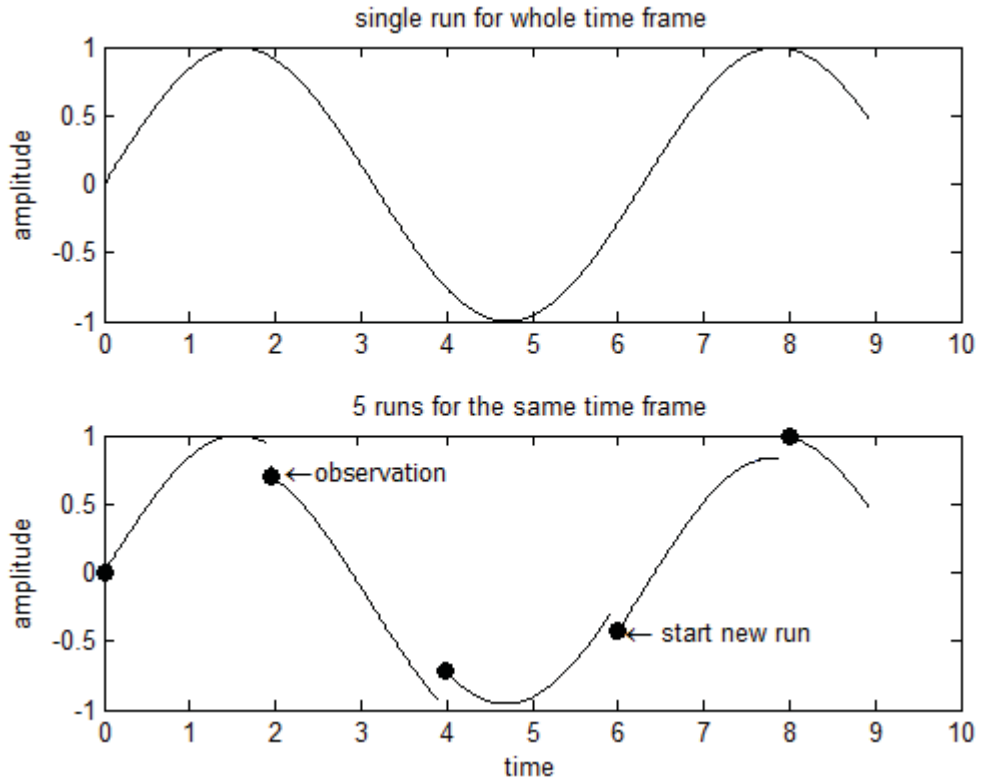


Figure 53: Difference between a normal simulation and a EnKF simulation for an arbitrary model

In order to demonstrate the compatibility of Sobek and OpenDA, the EnKF is first applied to a small two node hydraulic model. Subsequently, data is assimilated is applied to the case study.

7.1 Two node model with constant inflow

A schematisation of the simple two node model is presented in Figure 54. The hydraulic load consists of a constant discharge at one of the nodes.

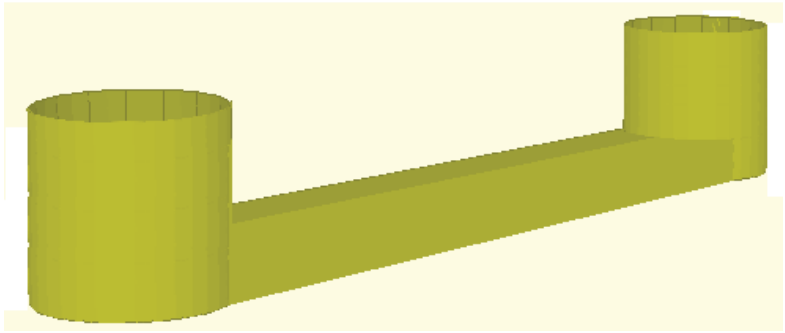


Figure 54: schematisation of the Sobek model containing one link and two computational nodes

In order to further simplify the model, the cross section of the pipe linking the two nodes is dimensioned large compared to the inflow. Therefore the system will react as a reservoir with an equal water level at both nodes. This means it is a valid assumption to neglect the momentum balance, and the state only needs to contain the water level. Water levels are obtained from the model output of the last time step, and imported in the initial condition file after the addition of noise by a customized java file.

The observations consist of perturbed water levels at the node opposite of the node with the inflow, derived from a copy of the model with discharge Q_1. Imperfections in the model are simulated by specifying a different flow rate in the assimilation model, namely Q_2. This means that even if the water level is corrected to approach the observation it will deviate again over time, making an update of the model state necessary when new observation are available. Robustness of the EnKF is tested further by specifying the wrong initial water level. The resulting water levels are presented in Figure 55.

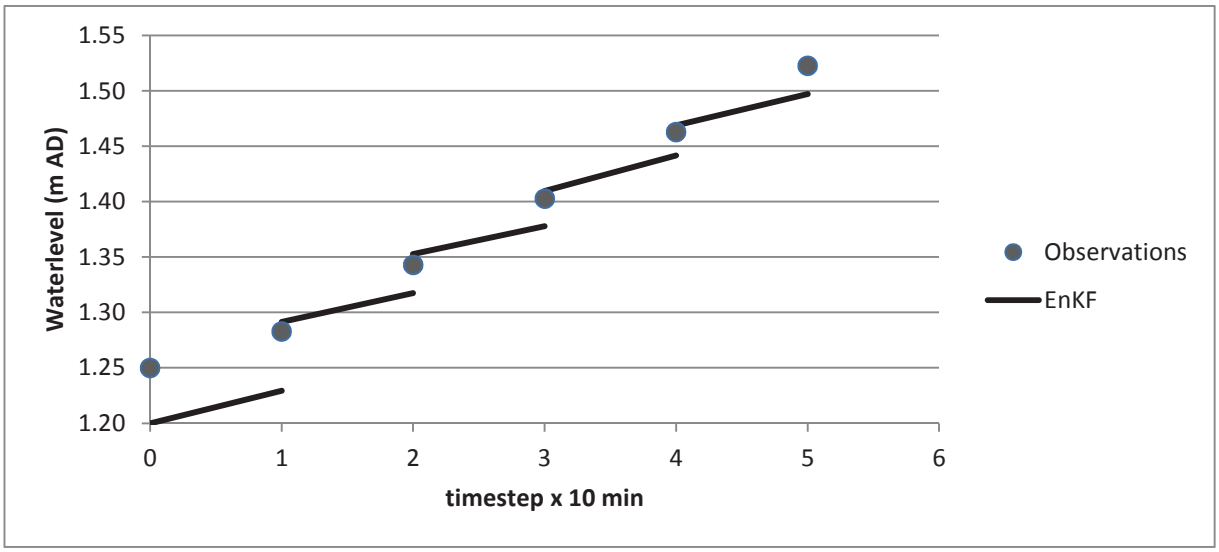


Figure 55: Observations and model predictions of the EnKF

The ensemble Kalman filter performs as expected; after one timestep the calculated water level equals the observation. Due to the fact that Q_2 is smaller than Q_1 the water level requires correction each time a new observation is available.

7.2 Delft city centre case study

Subsequently, the EnKF is applied to the case study, where the monitoring network designed in chapter 6 is used to provide the necessary observations. This monitoring network is designed using an integrated sewer and surface water model. However, the model used for data assimilation only contains the sewer model. This due to the following reasons:

- The surface water model has been constructed to run the channel flow module and the runoff module simultaneously in order to simulate interactions between the components, while OpenDA is programmed to run the modules in sequence.
- The necessary java files required to incorporate the boundary conditions into the state are not drafted at this time.

For both the runoff module and the sewer flow module, the state at the end of each sub run needs to be transferred in order to serve as an initial condition for the next sub run. Sobek has the ability to create restart files at the end of each run. A restart file contains initial conditions that can be used as a starting point for a new simulation. Noise needs to be added to the state of each ensemble member in order to introduce the spread in the model ensemble. Since no Java file is drafted to read or write these restart files at this time, OpenDA is not able to add noise to the complete state. This results in the setup is schematized in Figure 56. The state is transferred to serve as initial condition for the next sub run, but is not updated since no noise is added.

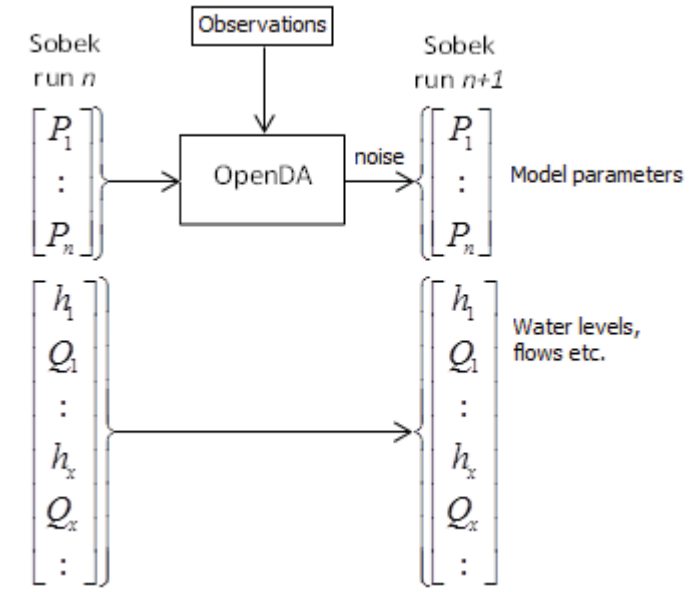


Figure 56: Setup used to transfer the state after each analysis step

This results in a situation where any deviation with the observations can only be compensated by a change in the parameter values, and the state is only influenced indirectly. This means that if the effect of a parameter is not or only partly perceptible within one analysis time step, the algorithm is inclined to change the parameter even further. This will eventually result in an overshoot of the model prediction. This principle is schematically shown in Figure 57.

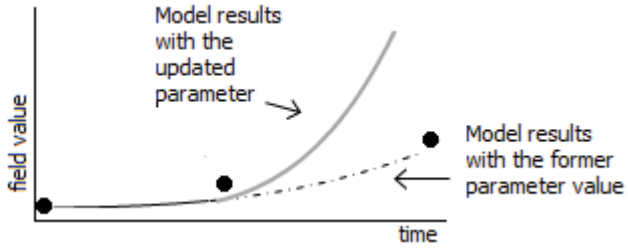


Figure 57: Divergence of the optimisation method due to

An example of such a parameter can be the runoff delay parameters in the runoff model. In order to prevent this phenomenon from occurring, noise should be added to the complete state of each ensemble member.

For the application of data assimilation for the case study, the observations are derived from water level series created by Sobek for the locations denoted as monitoring location in Chapter 6. The first run had an analysis step every 15 minutes, with a computational timestep of 1 minute and a total duration of 90 minutes. Evolution of the water depth for an arbitrary location is represented by the dotted grey line in Figure 58. Since no parameter was optimized, one would expect this line to be equal to water depth derived if the complete timeframe was run at once (black line). It can be seen that the six simulations of each 15 minutes do not only result in a delay in the water level peak , but also lower water levels in general. This suggests a discrepancy in the inflow in the sewer system coming from the runoff model.

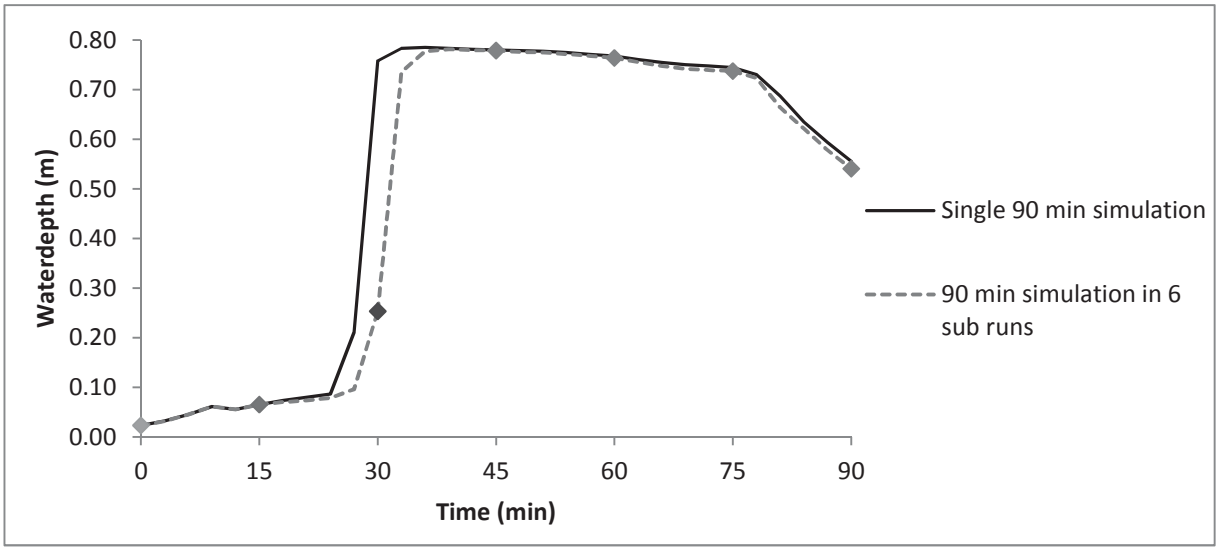


Figure 58: Difference in calculated water depth when the model is run in several parts

More insight in this phenomenon is gained by comparing the water balances of the runoff model. The precipitation for both simulations is presented in Figure 59. Over time the difference in the total amount of precipitation increases. Since the same water balance shows that the relation between the precipitation and the runoff is equal for both simulations, the difference in runoff can be attributed to a loss of precipitation.

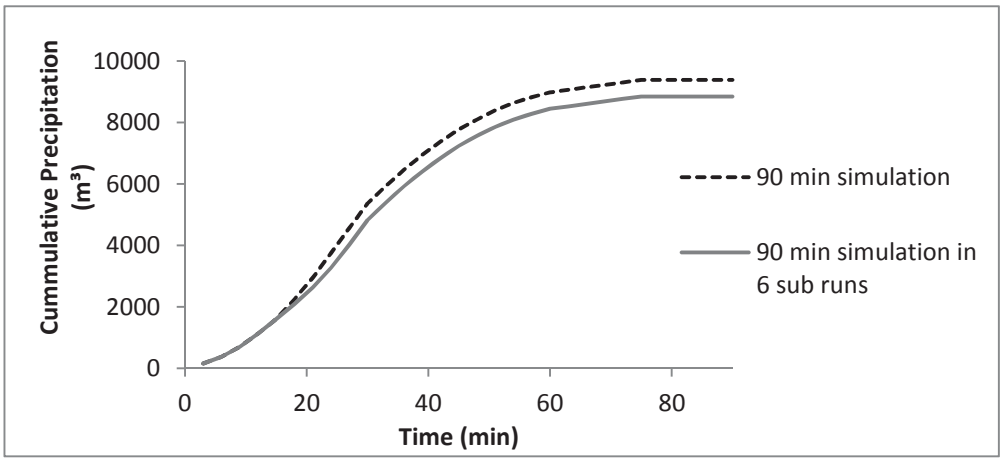


Figure 59: Cumulative precipitation for when the storm event is simulated in one run and in six sub runs

Decreasing the analysis time step and thereby increasing the number of sub runs needed to complete a simulation period of 90 minutes, results in an increasing amount of precipitation lost. Therefore it can be concluded that the smallest numerical error is obtained when the analysis time step is as large as possible. In order to decrease the loss of precipitation the timestep of the storm event is reduced from 5 minutes to 6 seconds. After this modification the difference between the single run and the 6 sub runs is negligible. However, this comes at a price; since the calculation time step cannot be larger than the storm event time step, the calculation time step needs to be reduced as well. This increases the computational load dramatically.

7.2.1 Analysis of the results

Results for a run with 100 ensemble members where the pipe roughness and the dry weather flow are optimized is presented in Figure 60. The EnKF does not perform as intended, since the simulated water depth deviates from the observations even though the initial parameter values were correct.

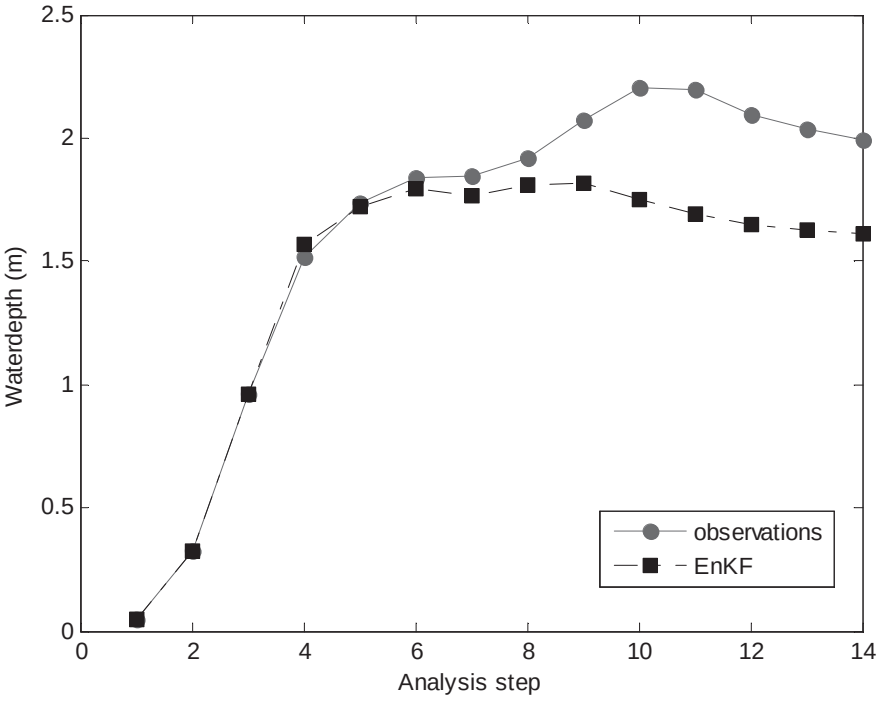


Figure 60: simulated water depth for one of the monitoring locations compared to the observations

After the third analysis step the EnKF slightly over estimates the water depth. This is compensated at the fifth analysis step, but results in parameter values that underestimate the water depth again the next step. At this point the estimated parameter values become negative, and therefore the model is unable to produce reliable results for the rest of the simulations. The lack of performance of the EnkF can likely be ascribed to the fact that no noise is added to the state, that for the flow model consists of the water levels and flows for the different calculation points. Although an increasing number of sub runs introduces a numerical error, the calculation time step of six seconds is deemed to be sufficient small in order to limit the effect on the model output.

The total computation time needed for a EnKF simulation with an ensemble size of one is comparable to the computation time needed for a normal simulation. It is observed that the computation time needed is proportional to the ensemble size as expected. Therefore the feasibility of the EnKF in the field of urban drainage is directly linked to the ensemble size needed to produce reliable results.

Chapter Summary

- The EnKF is implemented in Sobek by OpenDA. Communication between the model and OpenDA takes place through several model specific Java files.
- For a simple 2-node model the EnKF performs as expected. The filter has no problem with the initial estimation or the incorrectly specified flow.
- The performance of the EnKF for the Case study is poor. This can likely be ascribed to the fact that at this point, not all the Java files needed to add noise to the complete state for this particular model are drafted.
- A balance error in the precipitation is observed when the EnKF is implemented in OpenDA. Although the cause of this error is not found, a smaller model time step does result in a reduction of this error at the cost of an increase in computational load.

8 Conclusions and recommendations

The basis of this thesis is the application of data assimilation in the field of urban drainage and the design of a monitoring network. The Ensemble Kalman Filter (EnKF) is introduced as data assimilation method and tested for a simple example implemented in Matlab®. Furthermore, the EnKF is implemented by the open source toolbox OpenDA for the hydrodynamic software package Sobek.

Theory on the optimisation algorithm for the design of a monitoring network as proposed by (Henckens & Clemens, 2004) is expanded. The information content of potential monitoring locations is judged by analysing the singular values of the Jacobian matrix, which are calculated using a singular value decomposition. Furthermore, the influence of the sampling interval on the information content is evaluated in order to derive a measuring frequency.

8.1 Conclusions

The conclusions are divided into a general section, and a case specific section.

general

- In addition to the ability to identify relevant model parameters, a singular value decomposition can also be used to judge the information that potential monitoring locations can collect on these parameters.
- applying the expanded de-correlation algorithm in the optimisation process results in a monitoring network where each sensor has at least one well correlated sensor. This makes the network less sensitive to the effects sensor failure, since the information loss in case of a faulty sensor will not be as severe. Moreover, the overlapping information can be used for the cross validation of the collected data.
- Since the number of combinations of locations increases extremely when the total number of potential monitoring locations is expanded, it is not feasible for larger networks to calculate all possible combinations in order to find the best set of monitoring locations. Applying a genetic algorithm is a practical alternative, capable of finding a set of locations that approaches the combination that provides the most information on the system. A sufficient large population is a prerequisite for the success of this algorithm. The population size needed is depended on the number of locations where a sensor can be installed and the number of sensors to be placed.

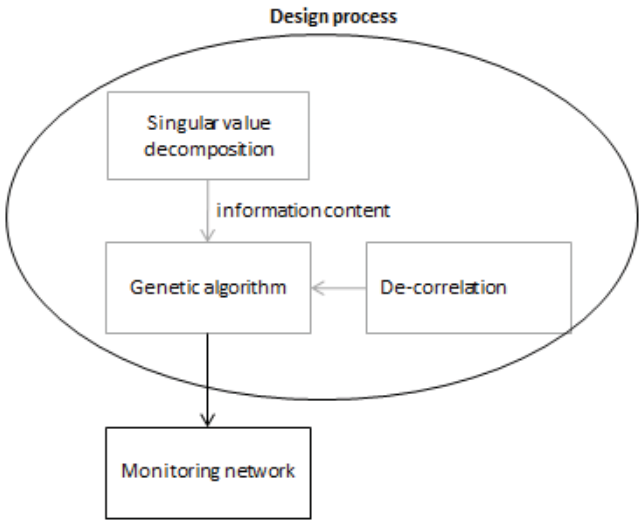


Figure 61: abbreviated schematisation of the optimisation process

- The optimisation process schematized in Figure 61 is an abbreviated version of Figure 26, and shows the relation between the prior enumerations in this paragraph. This optimisation process results in a set of monitoring locations that maximises the information content for a fixed number of sensors and is able to provide sufficient data for data assimilation.
- Since the EnKF integrates an ensemble of states forward in time in order to estimate the process noise covariance matrix, a larger ensemble size will result in a better estimation. Since the computation time needed is approximately proportional to the ensemble size, the feasibility of the EnKF in the field of urban drainage will depend on the ensemble size needed for the system in question.
- A set of parameter values belonging to a reservoir model have successfully been determined by implementing the EnKF as data assimilation method in Matlab®. The EnKF proves to be robust with respect to the initial estimation of the parameter values and the variability in the model input.

Case specific

- Relevant process parameters for the case study are those related to the run-off model, dry weather flow, pipe friction and most weir coefficients. Some weir coefficients can be omitted due to the fact that they correspond to nearly zero singular values, others can be clustered because they cannot be identified separately.
- Friction in the surface water system does influence the sewer system, indicating an interaction between both systems. However, this influence is limited indicating the surface water system created by closing the adjustable weirs reacts more like a reservoir for the storm events applied.
- Analysis of the correlation matrix indicates that the de-correlation algorithm is successfully applied; most sensors have one well correlated sensor, while the correlation is minimized for all other locations.
- Increasing the sampling interval, results in a relative steady decrease of the singular values for the case study. However, for a sampling interval larger then 3 minutes it is found that it is no longer possible to separately identify most of the weir coefficients. Therefore one can conclude that based on the results of a singular value decomposition an upper boundary for the sampling interval can be determined.
- OpenDA has been successfully used to implement the EnKF for the hydrodynamic software package Sobek. Predictions for a simple model consisting of one link and two computational nodes improved significantly. Tests with a sewer model comprising the city centre of Delft have been unsuccessful at this point. Due to time constraints, the necessary java files needed to add noise directly to the state are not compiled for this case. This is likely to be the cause of the poor performance for this particular case. Therefore, at this time it cannot be concluded that data assimilation can be applied to models in urban drainage in order to simulate field observations for continuous time series.
- For the example with the simple Sobek model, an ensemble size of 30 is sufficient in order to obtain reliable estimations. However, this example is not considered to be representative for the complex sewer systems found in practice. Therefore no confirmation can be made about the ensemble size needed for the case study.

8.2 Recommendations

The recommendations are divided in a section concerning the design of a monitoring network and a section concerning data assimilation.

Data assimilation

- In its current form, noise is not added to the complete state of the Sobek model, likely resulting in the poor performance of the EnKF for the case study. Therefore it is recommended that the

necessary Java files are drafted so that the application of the EnKF for the case study can be studied.

- Further research is needed on the effect of an increasing number of ensemble members on the accuracy of the EnKF estimation for hydrodynamic models in urban drainage in order to demonstrate the feasibility. (Gillijns, et al., 2006) propose to use the mean squared error to evaluate the effect of the ensemble size on the accuracy of the results.
- The EnKF splits a simulation up in a number of simulations with a shorter timeframe in order to implement improvements to model parameters. Running several simulations for a shorter timeframe introduces a balance error inherent to the model used, due to the fact that a part of the precipitation does not end up in runoff model. decreasing the storm event time step reduces this error, but increases the computational load. This balance error should be investigated more thorough in order to find a less computational expensive solution.
- In its current form, the implementation of the EnKF for Sobek requires the drafting of separate java files for almost each parameter or quantity of interest. Therefore it is advised to change the Sobek database structure to be more compact so that fewer java files need to be drafted, and that the interaction between the different components of the model are more clear.

Design of a monitoring network

- The system in the case study used to test the de-correlation algorithm is considered to be a flat system. Since the correlation between monitoring locations is influenced by the slope of the system, the performance of this algorithm should also be investigated for a sloped system.
- Due to the fact that one storm event is used for the entire system, every part of the system experiences the same hydraulic load at the same time. This can result in spurious correlation between locations that may not even be in the same system (Henckens, 2012). It is advised to further research the occurrence of spurious correlation in order to secure the added value of correlation between sensors in practice.

List of references

Aanmondt, A. & Nygard, M., 1995. Different roles and mutual dependencies of data, information, and knowledge. *Data and Knowledge Engineering,* Volume 16, pp. 191 - 222.

Arfken, G., 1985. *Mathematical methods for physicists.* 3rd ed. Orlando, Florida: Academic Press.

Brubaker, M., 2006. *CSC320 Tutorial notes (MIT),* Cambridge, Massachusetts: Massachusetts Institute of Technology.

Burgers, G., van Leeuwen, P. & Evensen, G., 1998. Analysis scheme in the ensemble Kalman filter. *Monthly Weather Review,* Volume 126, pp. 1719 -1724.

Butz, T., 2006. *Fourier transformations for pedestrians.* 1st ed. Leipzig, Germany: Birkhäuser.

Clemens, F., 2001. *Hydrodynamic models in urban drainage: application and calibration.* Delft: Delft University Press Science.

Clemens, F., 2002. *Evaluation of a method for the design of monitoring networks in urban drainage..* Portland, Oragon, Ninth International Conference on Urban Drainage.

Clemens, F., Langeveld, J., Korving, J. & Henckens, G., 2005. *Calibration of hydrodynamic models in urban drainage: directing model improvements.* Delft: Delft University of Technology.

De Jong, K., 1975. *An analysis of the behavior of a class of genetic adaptive systems,* Ann Harbor, Michigan: University of Michigan Press.

Deep, K., Singh, K., Kansal, M. & Mohan, C., 2009. A real coded genetic algorithm for solving integer and mixed integer optimization problems. *Applied Mathematics and Computation,* 212(2), pp. 505 - 518.

Di Pierro, F. et al., 2005. Automatic calibration of urban drainage model using a novel multi-objective genetic algorithm. *Water science and technology,* 52(5), pp. 43 - 52.

Dijk, van, Z., 2005. *Gracht, Delft.* sl:Wikimedia Commons.

Drécourt, J., 2004. *Data assimilation in hydrological modelling,* Hørsholm, Denmark: Technical University of Denmark.

Eskes, H. et al., 1998. Variational data assimilation: How to extract more information from GOME total ozone data. *Earth Observation Quarterly ,* Issue 58, pp. 35 - 38.

Evensen, G., 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research,* 99(C5), pp. 10143 - 10162.

Evensen, G., 2003. The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dynamics,* Issue 53, pp. 343 - 367.

Evensen, G., 2009. *Data assimilation: the ensemble Kalman filter.* 2nd red. Berlin, DE: Springer.

Gershenfeld, N., 1999. *The nature of mathematical modelling.* 1st red. Cambridge, UK: Cambridge University Press.

Ghil, M. & Malanotte-Rizzoli, P., 1991. Data assimilation in meteorology and oceanography. *Advanced Geophysics,* Issue 33, pp. 141 - 266.

Gillijns, S. et al., 2006. *What is the ensemble Kalman filter and how well does it work?.* Minneapolis, Minnesota, American Control Conference.

Hager, W., 2010. *Wastewater hydraulics: theory and practice.* 2nd red. Zürich, Switserland: Springer.

Hamilton, J., 1994. *Time series analysis.* 1st red. Princeton, New Jersey: Princeton University Press.

Harder, R., 2010. *Data validation in environmental sensor networks,* Delft: Delft University of Technology.

Haupt, R. & Haupt, S., 2004. *Practical genetic algorithms.* 2nd red. Hoboken, New Jersey: Wiley-Interscience.

Henckens, G., 2003. Calibration of the hydrodynamic sewer model of Loenen (In Dutch), org title: Kalibratie van het hydrodynamische rioleringsmodel van Loenen. *Rioleringswetenschap en -techniek,* 3(11), pp. 49 - 60.

Henckens, G., 2012. *de-correlation* [Interview] (27 july 2012).

Henckens, G. & Clemens, F., 2004. *Design and optimisation of monitoring networks in urban drainage.* Meaux-la-Montagne, France, CityNet 19th European Junior Scientist Workshop on Process Data and Integrated Urban Water Modelling , pp. 1-8.

Henckens, G., Clemens, F. & Stigter, L., 2007. Time series calibration of hydrodynamics models in sewer systems. *Sewer processes and networks ,* pp. 101-110.

Henckens, G., Schilperoort, R. & Clemens, F., 2005. *Monitoring network design using multiple storm events.* Copenhagen, Denmark, Proceedings of the 10th International Conference on Urban Drainage , pp. 1-10.

Hénonin, J. et al., 2010. *Urban flood real-time forecasting and modelling: A state-of-the-art review.* Copenhagen, sn, pp. 1-21.

Henrichs, M., Vosswinkel, N. & Uhl, M., 2008. *Influence of uncertainties on calibration results of a hydrological model.* Edinburgh, Scotland, 11th International Conference on Urban Drainage, pp. 1-10.

Holland, J., 1975. *Adaptation in natural and artificial systems.* Ann Arbor, Michigan, University of Michigan Press.

Hutton, C., Vamvakeridou-Lyroudia, L., Kapelan, Z. & Savic, D., 2010. *Real-time modelling and data assimilation techniques for improving the accuracy of model,* s.l.: Seventh Framework Programme.

Johnson, C., 2003. *Information content of observations in variational data assimilation,* Reading, England: University of Reading.

Kalman, D., 1996. A Singularly valuable decomposition: The SVD of a matrix. *The College Mathematics Journal,* 27(1), pp. 2-23.

Kalman, R., 1960. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering,* 82(series D), pp. 35 - 45.

Kalman, R. & Bucy, R., 1961. New results in linear filter and prediction theory. *Journal of Basic Engineering,* 83(1961), pp. 95 - 108.

Kang, D. & Lansey, K., 2009. Real-time demand estimation and confidence limit analysis for water distribution systems. *Journal of Hydraulic Engineering,* 135(10), pp. 825 - 837.

Kleidorfer, M., Deletic, A., Fletcher, T. & Rauch, W., 2009. Impact of input data uncertainties on urban stormwater model parameters. *Water Science and Technology,* 60(6), pp. 1545 - 1554.

Kleidorfer, M., Möderl, M., Fach, S. & Rauch, W., 2009. Optimization of measurement campaigns for calibration of a hydrological model. *Water Science & Technology,* 59(8), pp. 1523 - 1530. .

Korres, G., Hoteit, I. & Triantafyllou, G., 2007. Data assimilation into a Princeton ocean model of the mediterranean sea using advanced Kalman filters. *Journal of Marine Systems,* 65(2007), pp. 84 - 104.

Korving , J., 2012. *Data-assimilation in Intelligent measuring (In Dutch), org. title: Data assimilatie in Slim meten,* Delft: Technical University Delft.

Korving, J., van Gelder, P., van Noortwijk, J. & Clemens, F., 2002. *Influence of model parameter uncertainties on decision making for sewer system management.* Cardiff, England, Proceedings of the 5th international conference on hydroinformatics Vol. 2 , pp. 1361-1366.

Langeveld, J., 2004. *Interactions within wastewater systems,* Delft: Delft University Press.

Langeveld, J., Clemens, F. & Henckens, G., 2004. *Integrated modelling and data needs: quantification based on the interactions within the wastewater system..* Meaux-la-Montagne, France, Proceedings of Citynet 19th European junior scientist workshop, pp. 1-14.

Lay, D., 2006. *Linear algebra and its application.* 3rd ed. Boston, Massachusetts: Pearson Education Inc.

Leeuwenburgh, O., 2005. *Implementation and testing of an ensemble Kalman filter assimilation system for the Max Planck Institute ocean general circulation model,* De Bilt: Koninklijk Nederlands Meteorologisch Instituut.

Lohuizen, van, C.W.W., 1986. Knowledge assessment and policy making. *Knowledge-Creation Diffusion Utilization,* 8(1), pp. 12-38.

Mandel, J., 2006. *Efficient implementation of the ensemble Kalman filter,* Denver, Colorado: University of Colorado at Denver and Health Sciences Center.

Marsalek, J., Maksimovic, C., Zeman, E. & Price, R., 1998. *Hydroinformatics tools for planning, design, operation and rehabilitation of sewer systems.* 2nd ed. Dordrecht: Kluwer Academic Publishers.

Mathworks, 2007. *How the genetic algorithm works.* [Online] Available at: http://www.mathworks.nl/help/toolbox/gads/f6187.html [Geopend 07 05 2012].

Mathworks, 2012. *Fast Fourier transform.* [Online] Available at: http://www.mathworks.nl/help/techdoc/ref/fft.html [Geopend 20 March 2012].

Municipality of Delft, 2012. *History of Delft.* [Online] Available at: http://www.delft.nl/Toeristen/Historie_van_Delft/Ontstaansgeschiedenis [Geopend 02 07 2012].

Nederlands Normalisatie Instituut, 2008. *NEN-EN 752:2008,* Delft: Nederlands Normalisatie-Instituut.

Nederlands Normalisatie Instituut, 1994. *Dutch practice outdoor sewerage management NPR 3220 (In Dutch), org. title: Nederlandse praktijkrichtlijn buitenriolering beheer,* Delft: Nederlands Normalisatie Instituut.

Nooyen, R. & van Overloop, P., 2008. *CT5490 Operational watermanagement,* Delft: Technical University Delft.

Olsthoorn, T., 1998. *Groundwater modelling: calibration and the use of spreadsheets.* Delft: Delft University Press.

Papadimitriou, C. & Steiglitz, K., 1998. *Combinatorial optimization.* 2nd red. Mineola, New York: Dover Publications inc..

Rauch, W. & Harremoës, P., 1999. On the potential of genetic algorithms in urban drainage modeling. *Urban Water,* 1(1), pp. 79 - 89.

Rauch, W. et al., 2011. *Uncertainty in online predictions of urban drainage models.* Porto Alegre, Brazil, 12nd International Conference on Urban Drainage, pp. 1-9.

Rioned, 2003. *Monitoring and calulating sewer systems (In Dutch), org. title: Meten en berekenen rioolstelsels,* Ede: Rioned Foundation.

Rioned, 2009. *Module C2330 measuring equipment (In Dutch), org. title: Module C2330 meetapparatuur,* Ede: Rioned Foundation.

Schilperoort, T., 1986. Statistical Aspects. In: *Design aspects of hydrological networks.* Delft: TNO Committee on Hydrological Research, pp. 35 - 56.

Seber, G. & Wild, C., 1989. *Nonlinear regression.* 1989 red. New York, New York: John Wiley & Sons.

Shi, Y. & van Albada, G., 2007. *Computational science -- ICCS 2007.* 1st red. Beijing, China: Springer.

Simon, D., 2001. Kalman filtering. *Embedded Systems Programming,* 14(6), pp. 72 - 79.

Solonen, A., 2011. *extended Kalman filter (EKF).* [Illustraties] (Lappeenranta University of Technology).

Sparknotes, 2012. *Applications of the derivative.* [Online] Available at: http://www.sparknotes.com/math/calcbc1/applicationsofthederivative/section2.rhtml [Geopend 07 05 2012].

Speed, D. & Ahlfeld, D., 1996. *Diagnosis of structural identifiability in groundwater flow and solute transport equations.* Golden, Colorado, IAHS publications.

Tait, S. & ten Veldhuis, J., 2011. *Data-driven urban drainage analysis: an alternative to hydrodynamic models?.* Porte Alegre, Brazil, 12nd International Conference on Urban Drainage, p. 9.

Ummels, T. & Clemens, F., 1998. *Propagation of errors in hydrodynamic calculations in urban drainage.* London, England, 4th international conference on developments in Urban Drainage Modelling, pp. 221-228.

UN/ECE Task Force on Monitoring & Assessment, 2000. *Guidelines on Monitoring and Assessment of Transboundary Rivers,* Lelystad, NL: RIZA.

Urbano , L., 2010. *Evolution and drug resistant bacteria.* [Online] Available at: http://montessorimuddle.org/2010/01/01/evolution-and-drug-resistant-bacteria/ [Geopend 07 05 2012].

Veldhuis, ten, J.A.E., 2010. *Quantitative risk analysis of urban flooding in lowlands areas,* Delft: Delft University of Technology.

Ven, van de, F.H.M., 2011. *Water management in urban areas,* Delft: Delft University of Technology.

Wang, B., Zou, X. & Zhu, J., 2000. Data assimilation and its applications. *Proceedings of the National Academy of Sciences of the United States of America,* 97(21), pp. 11143 - 11144.

Wijngaard, J. & Kok, M., 2004. *New rainfall statistics for water managers (In Dutch), org. title: Nieuwe neerslagstatistiek voor waterbeheerders,* Amersfoort: STOWA.

Annexe I: Singular values and Eigenvectors for the weirs

Singular values	Eigenvectors																																
0.00E+00	Weir number	3	8	13	14	20	21	22	23	42	46	49	15+16	17+19	24+25	29+33	36+38	31	34+35	5+6+7	4+11+9+10+18	30+32+37	39+41+43+26+44
4.18E-18		1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2.10E-17		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4.57E-17		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.06E-16		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
8.24E-02		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9.87E-02		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0
1.19E-01		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.85E-01		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.3	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.9	0.0	0.0
2.34E-01		0.0	0.0	0.0	0.0	0.0	0.1	0.0	1.0	-0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	0.0	0.0	0.0
2.95E-01		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	-0.6	-0.1	0.0	0.0	0.0	0.0	0.0	0.0	-0.8	0.0	0.0	0.0	0.1
2.99E-01		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	-0.8	-0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.1	0.0	-0.1
3.30E-01		0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	-0.9	0.0	0.0	0.0	-0.2	-0.1	0.0	0.0	0.0	0.2	0.0	-0.3	0.0	0.1
6.55E-01		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	-0.1	0.0	-0.9	0.0	0.2	0.0	0.0	0.3	0.0	0.1	0.2
6.99E-01		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	-1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7.61E-01		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	-0.2	-0.7	0.0	0.0	0.1	0.0	0.7	0.1
9.10E-01		0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	0.1	0.0	0.0	0.3	0.0	-0.2	-0.1	0.2	0.0	0.0	-0.9	0.0	0.3	0.2
1.07E+00		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	0.0	0.0	0.0	0.0	0.0	-0.2	0.4	0.0	-0.1	0.1	0.0	0.4	-0.8
1.53E+00		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	-0.4	0.4	-0.5	0.0	0.0	-0.2	0.0	-0.3	-0.5
2.57E+00		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8	0.0	0.1	-0.5	0.0	0.0	0.0	0.3	0.0	-0.1	0.1
3.32E+00		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.4	0.0	-0.1	-0.8	-0.2	0.0	0.0	-0.2	0.0	-0.4	-0.2
7.21E+00		0.0	-1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Annexe II: Applied storm events

The historical storm events used in this thesis are derived from a single rain gauge located near the south-east boundary of the case study. This device does not record the rain depth on a fixed time interval; instead the sampling frequency is increased depending on the intensity of the storm. In order to obtain a value for the rain intensity, differentiation and integration with fixed boundaries is applied.

Since the layout of a monitoring network is strongly influenced by the storm event applied (Henckens & Clemens, 2004), three different storms are used for the design in this thesis. It should be noted that the storm choice does not only influence the layout of the monitoring network, but also the determination of the relevant parameters due to the fact that the singular values are analysed for every storm. The total rain depth of each storm is sufficiently large in order to fill the entire sewer system and allow for the occurrence of CSO's.

The first storm selected is standard design storm 02 from the Dutch sewer guidelines, which corresponds to a return period of four times per year. This storm is relatively short and is characterised by a high peak at the end of the storm, as seen in Figure 62. The model calculates that pluvial flooding occurs locally when a design storm with a higher return period is used.

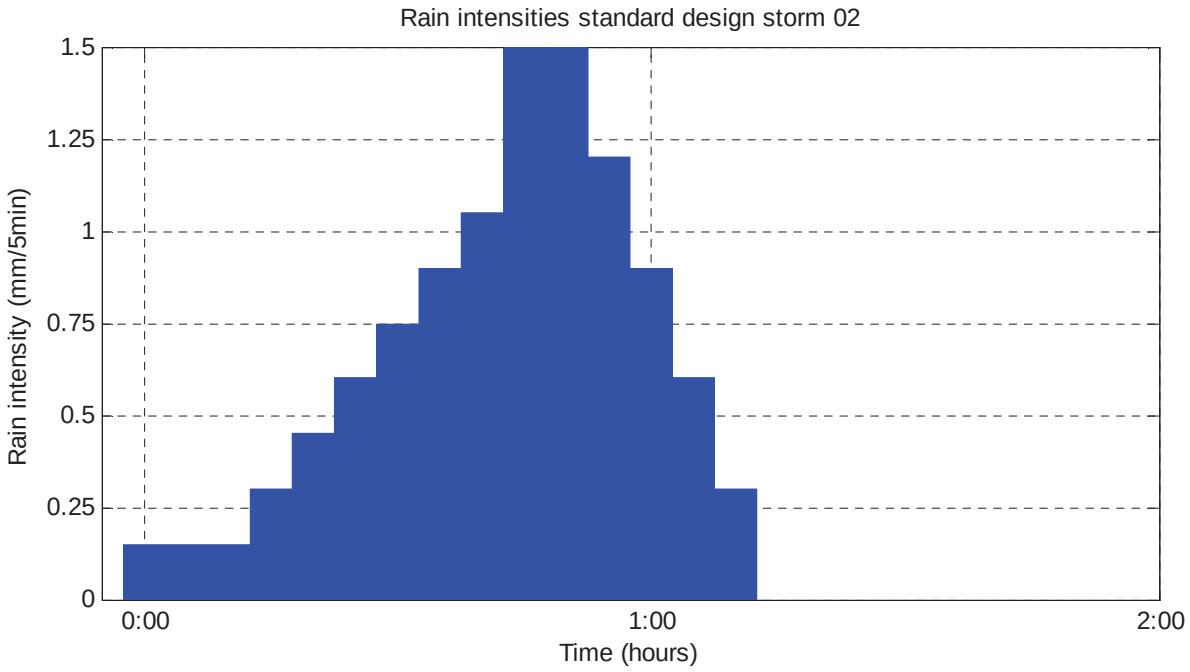


Figure 62: Standard design storm 02 with a return period of 0.25 years

The first historical storm is chosen from Table 7. This table comprises storm events with a daily precipitation sum between 15 and 25 mm. these somewhat arbitrary boundaries are stated to find a historical storm with a longer duration compared to design storm 02 without resulting in pluvial flooding. Based on the peak intensity, the location of the peak in the storm and a sufficient long dry period before and after the storm, the storm on 19-01-12 is chosen.

Table 7: Storm events with a precipitation sum between the 15 and 25 mm

Event number	Date	Daily sum (mm)	Peak intensity (mm/5min)
1	19-06-11	21.0	1.80
2	12-07-11	20.4	1.45
3	08-08-11	15.1	0.78
4	26-08-11	16.8	6.33
5	27-08-11	21.4	1.69
6	07-10-11	18.0	2.22
7	01-12-11	21.0	0.90
8	16-12-11	16.4	1.06
9	03-01-12	18.4	2.27
10	19-01-12	18.4	0.46
11	07-03-12	15.6	0.36
12	28-04-12	15.0	0.85
13	02-05-12	16.2	3.76
14	04-06-12	15.2	0.52
15	11-06-12	15.1	3.72
16	15-06-12	15.6	1.66

Progress of this storm is seen in Figure 63. The duration is approximately 12 hours, and has several peaks with a maximum intensity of 0.46 mm /5 min.

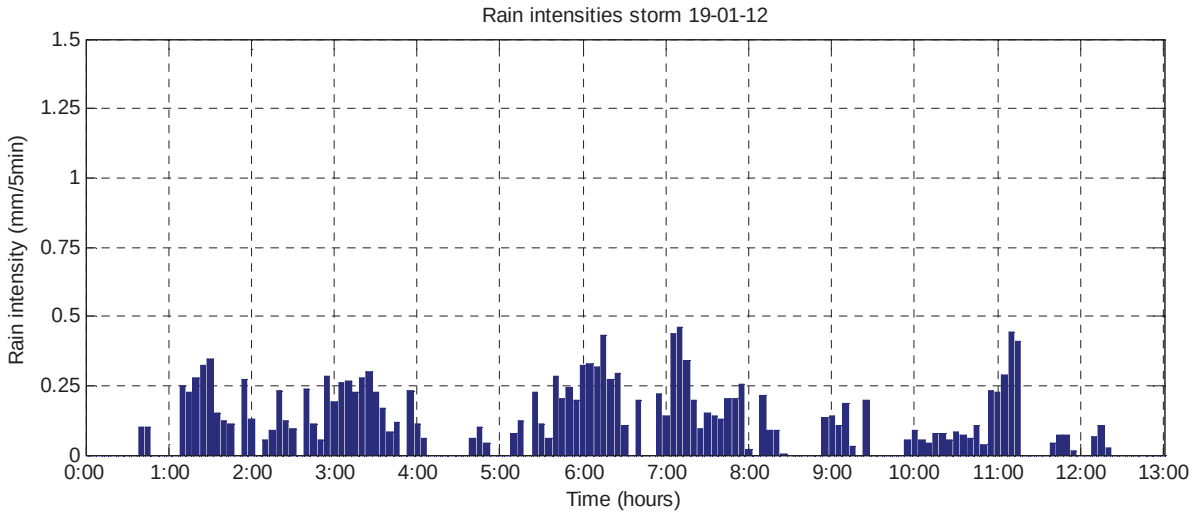


Figure 63: Historical storm 19-01-12

The second historical storm event is chosen to provide information on the importance of surface water related parameters. According to (Wijngaard & Kok, 2004)return periods for water systems varying between the 0.5 years and two years correspond to a daily precipitation amount between 28 and 39 mm. The only storm event to meet this criteria in the time series available, is the storm of 24-07-11. The total amount of precipitation is 37 mm with a peak of 0.95 mm/5 min.

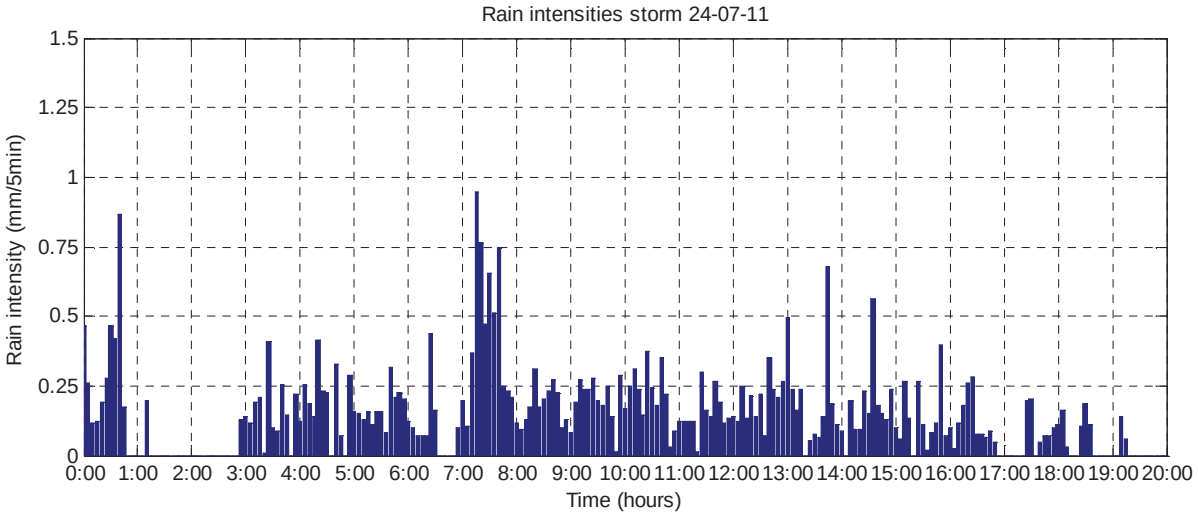
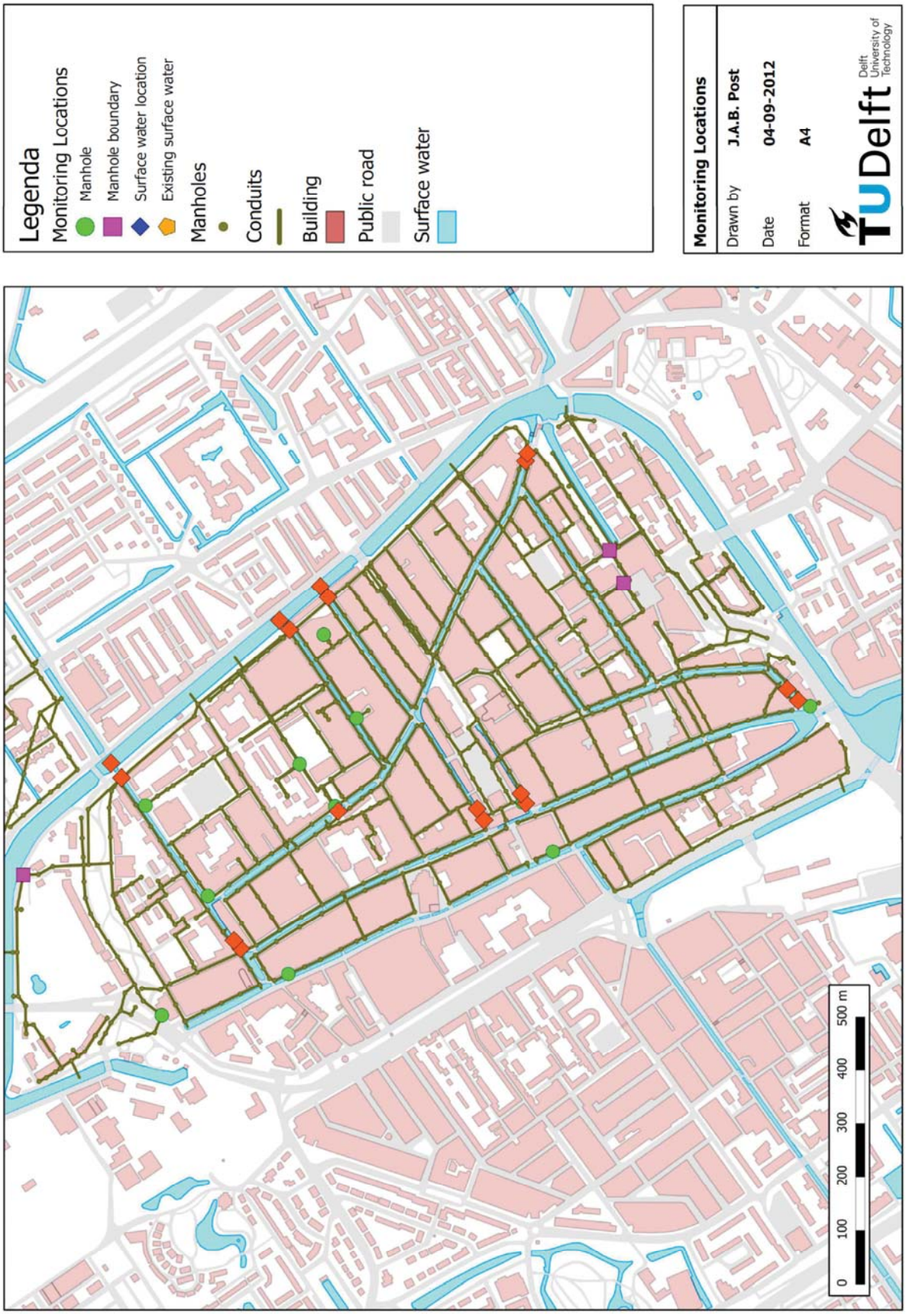


Figure 64: Historical storm 24-07-11

Annexe III: Singular values and Eigenvectors for the applied storms

Parameters	3.57E+04	8.40E+02	3.86E+02	2.69E+02	2.12E+02	1.58E+02	1.13E+02	1.09E+02	8.62E+01	7.66E+01	7.17E+01	5.34E+01	4.69E+01	4.22E+01	3.60E+01	3.00E+01	2.22E+01	1.28E+01	9.19E+00	8.51E+00	8.14E+00	7.47E+00	6.73E+00	4.96E+00	3.81E+00	4.22E-01	3.03E-01
	Singular values																										
	Eigenvector																										
R1	0.0	-0.3	0.0	0.9	-0.1	0.2	0.0	0.0	-0.1	0.0	-0.1	-0.1	0.0	0.0	-0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
R2	0.0	-0.1	-1.0	-0.1	-0.2	0.0	0.0	0.0	0.0	-0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
R3	0.0	-0.1	-0.1	0.1	0.1	-0.6	0.5	0.0	0.3	0.5	-0.2	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
R4	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	0.0	0.0	0.0	0.0	0.3	0.5	0.8	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
B1	0.0	0.0	0.0	0.1	0.1	-0.4	-0.6	0.0	-0.1	0.1	-0.3	0.0	-0.5	0.2	0.1	0.1	-0.2	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
B2	0.0	0.0	0.0	0.1	0.5	-0.2	0.3	-0.4	-0.6	-0.3	0.3	0.1	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
B3	0.0	-0.9	0.2	-0.3	0.0	0.0	0.0	-0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
B4	0.0	0.0	0.0	0.0	0.0	-0.1	-0.1	0.0	0.0	0.1	0.0	0.1	0.0	0.0	-0.2	0.3	0.0	0.0	-0.6	-0.5	-0.2	0.2	-0.4	-0.1	0.0	0.0	0.0
I1	0.0	0.0	-0.2	-0.1	0.8	0.4	0.0	0.3	0.0	0.3	-0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
D1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	-0.1	0.4	-0.7	-0.1	-0.4	0.2	0.0	0.0	0.0	0.0
W1	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	0.1	0.0	0.2	0.5	-0.2	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	0.3	0.3	-0.5	-0.5	0.0	0.0
W2	0.0	0.0	-0.1	0.1	0.3	-0.1	-0.2	-0.2	0.7	-0.5	0.2	-0.1	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
W3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	0.0	0.0	0.0	0.1	0.0	-0.5	-0.3	-0.2	0.6	0.1	0.5	0.1	0.1	0.0	0.0
W4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
W5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	0.0	0.0	0.0	0.0	-0.1	-0.8	0.0	0.2	-0.5	-0.2	-0.1	-0.1	-0.1	0.0	0.0
W6	0.0	0.0	0.0	0.0	0.0	-0.2	-0.2	0.1	0.0	0.2	0.1	0.3	0.2	-0.2	-0.5	0.6	0.1	0.0	0.2	0.2	0.1	-0.1	0.1	0.0	0.0	0.0	0.0
W7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.4	-0.4	-0.4	-0.7	0.2	0.0	0.0
W8	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	0.1	0.0	0.2	0.5	-0.2	0.0	0.1	0.0	-0.1	0.0	0.0	0.0	0.0	-0.2	0.1	0.1	0.1	0.8	0.0	0.0
W9	0.0	0.0	0.0	0.0	0.1	-0.2	-0.3	0.0	0.0	0.0	0.0	0.1	0.1	-0.1	-0.1	-0.5	0.7	-0.2	0.1	-0.1	0.0	0.1	-0.1	0.0	0.0	0.0	0.0
W10	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	0.1	0.0	0.2	0.5	-0.2	0.0	0.1	0.0	0.0	0.0	0.0	-0.1	0.0	0.2	-0.4	-0.4	0.5	-0.3	0.0	0.0
W11	0.0	-0.1	0.1	0.0	0.0	-0.2	0.2	0.8	0.0	-0.4	0.1	0.1	-0.3	0.2	-0.2	-0.1	0.0	0.0	-0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
W12	0.0	0.0	0.0	0.0	0.1	-0.3	-0.1	0.2	-0.2	-0.1	-0.2	-0.7	0.4	-0.1	0.2	0.2	0.1	0.1	-0.1	0.0	-0.1	-0.1	0.0	0.0	0.0	0.0	0.0
W13	0.0	-0.1	0.0	0.1	0.0	0.0	-0.1	0.3	0.0	0.0	0.1	0.4	0.1	-0.3	0.7	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
W14	0.0	0.0	0.0	0.0	0.1	-0.2	-0.2	0.1	0.0	0.0	0.0	0.1	0.3	-0.3	-0.2	-0.5	-0.6	0.1	-0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
F1	-1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
F2	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	-0.1	0.0	0.1	0.0	0.1	-0.1	0.0	0.0	-0.1	0.2	0.3	-0.5	0.2	-0.2	-0.6	0.4	0.0	-0.1	0.0	0.0
F3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0

Annexe IV: Layout of the monitoring network



Annexe V: Kalman filter example

The following example concerns with the correction of a navigation system and can, to a large extend, be regarded as a citation from (Simon, 2001). Because of the simple nature of the system it is a perfect case to provide insight in the propagation of the covariance matrices in time and the contents of the state equations.

First the State equation is described by

$$\underline{x}_{k+1} = \underline{\underline{A}}\underline{x}_k + \underline{w}_k$$

Where:

$\underline{x}_k$ = state at time k

$\underline{\underline{A}}$ = matrix containing information on the quantities affecting the state

$\underline{w}_k$ = process noise

The observations are related to the state by

$$\underline{y}_k = \underline{\underline{C}}\underline{x}_k + \underline{z}_k$$

Where:

$\underline{y}_k$ = observations at time k

$\underline{x}_k$ = state at time k

$\underline{\underline{C}}$ = matrix relating the state to the observations

$\underline{z}_k$ = measurement error

The process noise and measurement error are drawn from a normal distribution with mean zero and standard deviation σ_w and σ_z respectively. For the navigation system the state consists of the position and the velocity, while the acceleration and the time are known input. From the equation of motion we know that:

$$v_{k+1} = v_k + a_k \Delta t$$

Where:

v = velocity

a = acceleration

Δt = timestep size

Integration with respect to time yields

$$s_{k+1} = s_k + v_k \Delta t + \frac{1}{2} a_k \Delta t^2$$

Where:

s = position relative to s_0

v = velocity

a = acceleration

Δt = timestep size

These equations can be rewritten as a set of linear equations fitting the state and observation equations. The only quantity observed is the position.

$$\underline{x}_{k+1} = \begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix} \underline{x}_k + \begin{bmatrix} 0.5\Delta t^2 \\ \Delta t \end{bmatrix} \underline{a}_k + \underline{w}_k$$

$$\underline{y}_{k+1} = \begin{bmatrix} 1 & 0 \end{bmatrix} \underline{x}_k + \underline{z}_k$$

The acceleration is not placed in the state matrix since it is considered to be input by the driver of the vehicle. The measurement error covariance matrix and the process noise covariance matrix are defined as

$$\underline{\underline{S}}_w = E\left(\underline{w}_k \underline{w}_k^T\right)$$

$$\underline{\underline{S}}_z = E\left(\underline{z}_k \underline{z}_k^T\right)$$

Where:

$\underline{\underline{S}}$ = noise/error covariance matrix

E = expected value

T = transpose

$\underline{z}_k$ = measurement error

$\underline{w}_k$ = process noise

In order to derive the covariance matrices some assumptions concerning the example have to be made. It is assumed that the position is measured with an error of 10 meters (measurement error) and that commanded acceleration is a constant 1 m/s^2 with a process noise of 0.2 m/s^2. With a timestep of t = 0.1 s the following matrices are obtained

$$\underline{x}_{k+1} = \begin{bmatrix} 1 & 0.1 \\ 0 & 1 \end{bmatrix} \underline{x}_k + \begin{bmatrix} 0.005 \\ 0.1 \end{bmatrix} \underline{a}_k + \underline{w}_k$$

$$\underline{y}_{k+1} = \begin{bmatrix} 1 & 0 \end{bmatrix} \underline{x}_k + \underline{z}_k$$

In the model the error of the position is proportional to $0.5 \cdot \Delta t^2 \cdot a_k$ and the velocity to $\Delta t \cdot a_k$, with a_k = 0.2 the position and the velocity become proportional to $5 \cdot 10^{-3}$ and $2 \cdot 10^{-2}$ respectively. This relation is used to set up the following covariance matrix

$$\underline{\underline{S}}_w = E\left(\underline{w}\ \underline{w}^T\right) = E\left(\begin{bmatrix} s \\ v \end{bmatrix} \begin{bmatrix} s & v \end{bmatrix}\right) = E\begin{bmatrix} s^2 & sv \\ vs & v^2 \end{bmatrix} = \begin{bmatrix} \frac{1}{4}\Delta t^4 \cdot 0.2^2 & \frac{1}{2}\Delta t^3 \cdot 0.2^2 \\ \frac{1}{2}\Delta t^3 \cdot 0.2^2 & \Delta t^2 \cdot 0.2^2 \end{bmatrix}$$

Since only the measured position is subject to a measurement error, the error covariance is described by:

$$S_z = E\left(zz^T\right) = 10^2$$

The following Kalman filter equations are used to update the state:

$$\underline{x}_k^a = \underline{x}_k^f + \underline{\underline{K}}_k \left(\underline{y}_k - \underline{y}_k^f \right)$$

$$\underline{\underline{K}}_k = \underline{\underline{P}}_k \underline{\underline{C}}^T \left(\underline{\underline{C}} \underline{\underline{P}}_k \underline{\underline{C}}^T + S_z \right)^{-1}$$

$$\underline{\underline{P}}_{k+1} = \underline{\underline{A}} \underline{\underline{P}}_k \underline{\underline{A}}^T + \underline{\underline{S}}_w - \underline{\underline{A}} \underline{\underline{P}}_k \underline{\underline{C}}^T S_z^{-1} \underline{\underline{C}} \underline{\underline{P}}_k \underline{\underline{A}}^T$$

Where:

$\underline{x}_k^a$ = actual state

$\underline{x}_k^f$ = forecasted state

$\underline{\underline{K}}_k$ = Kalman gain

$\underline{y}_k$ = observation

$\underline{y}_k^f$ = prediction of the quantity to be observed

$\underline{\underline{P}}$ = model state error covariance matrix

$\underline{\underline{C}}$ = matrix relating the state to the observation

$\underline{\underline{A}}$ = matrix containing information on the quantities affecting the state

$\underline{\underline{S}}_w$ = process noise covariance matrix

S_z = measurement error covariance

Since the position and the velocity at the start are zero we know the initial condition as well as the initial error covariance matrix; if the state is exactly known the error covariance matrix is the zero matrix.

Annexe VI: OpenDA structure

The three main components specified in the main OpenDA file are the algorithm, stochastic observer and the stochastic model. The former contains the settings concerning the calibration or data assimilation method used. The stochastic observer refers to the observations used. The latter contains information on the model and the parameters or state and consists of three layers, each represented by one file. This scheme is presented in the diagram in Figure 65.

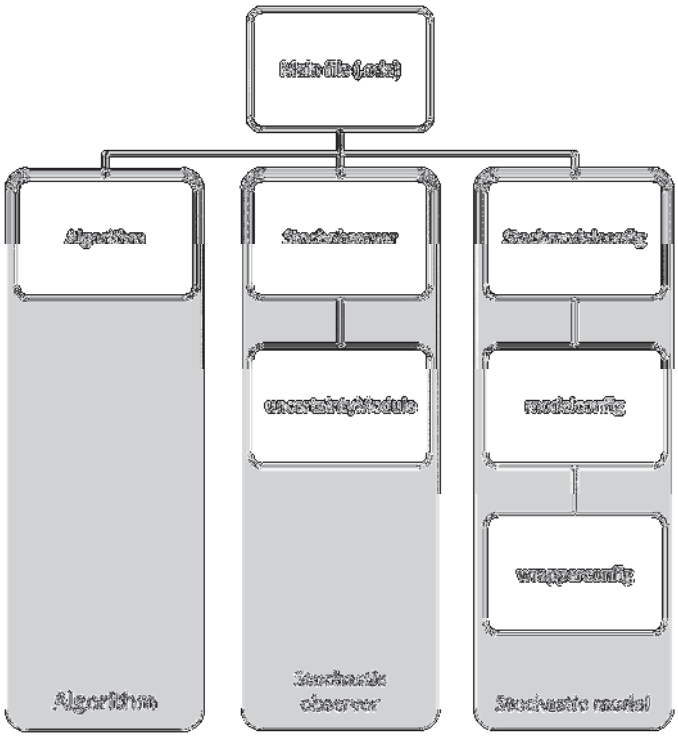


Figure 65: diagram of the xml build of OpenDA

Stochastic model

The schemes presented in Annexe VII show the relation between the xml files that comprise the Stochastic model, the corresponding java files and the model input for an arbitrary simple calibration example. Calibrating the constant dry weather flow in Sobek requires modifying the input file: "*pluvius.dwa*". The amount of DWF (l/hr*capita) is defined in the field next to the field containing the string *wc*. StochModelConfig.xml is the outer layer file and contains information on the parameters being calibrated and step sizes used in the process. Modelconfig is the middle layer and includes more detailed information on the location of the files being edited and the items being exchanged between the model and OpenDA. Finally, wrapperConfig is the inner layer contains references to the java files controlling the input files and information on the executables of the model. The java files are used to read and write the parameter values as OpenDA pleases.

Running the EnKF in OpenDA requires a somewhat different setup. This is due to the fact that an update is performed after each iteration and noise needs to be added to the state. Therefore instead of running a simulation based on the whole time series and then making another run with a different set of parameter values, each ensemble member is run one step. After the analysis, the state is updated and another simulation step is performed. It can occur that the file where the state is read from, is not the file used as initial condition for the next run. For the example presented in Annexe VII the dry weather flow is used, which can be accessed for output as well as input.

Stochastic observer

When no "real" measurements are available, simulation output can also be imported. Observations can be produced by a model for a specified measuring interval and locations. StochObserver.xml contains information on the java file needed to translate the observations to a standard comprehensible by OpenDA. To incorporate the noise present in real measurements, an extra noise term is specified in the uncertaintyModule. An example of the uncertainty file in the OpenDA interface is presented in Figure 66.

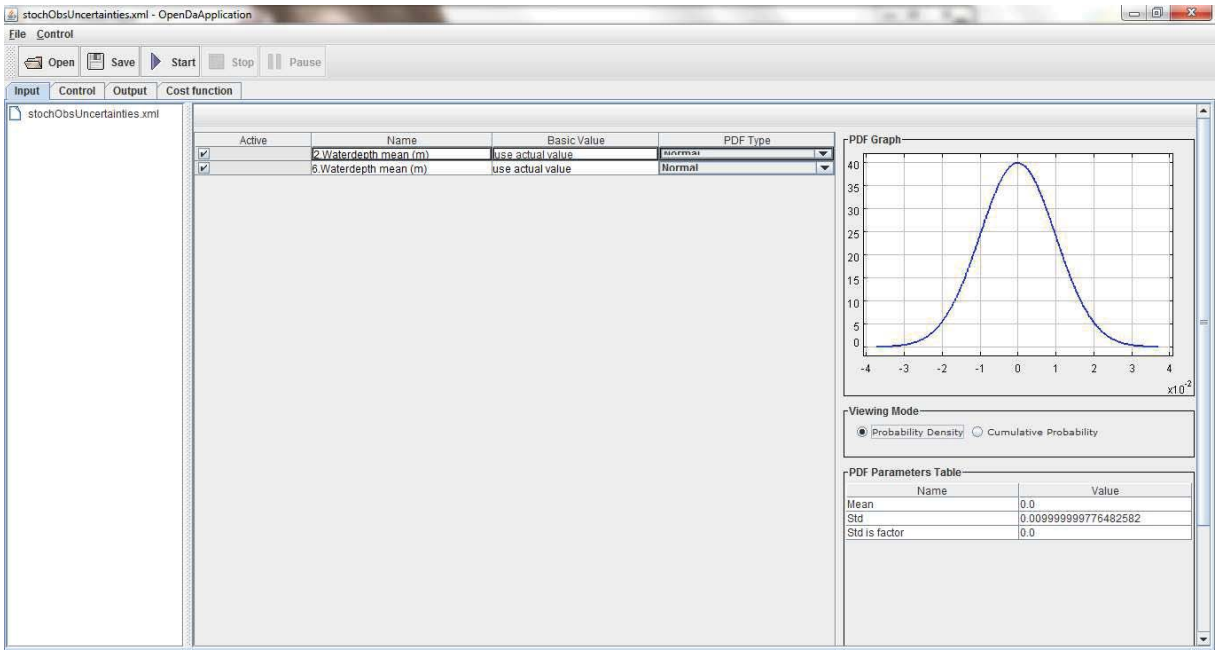
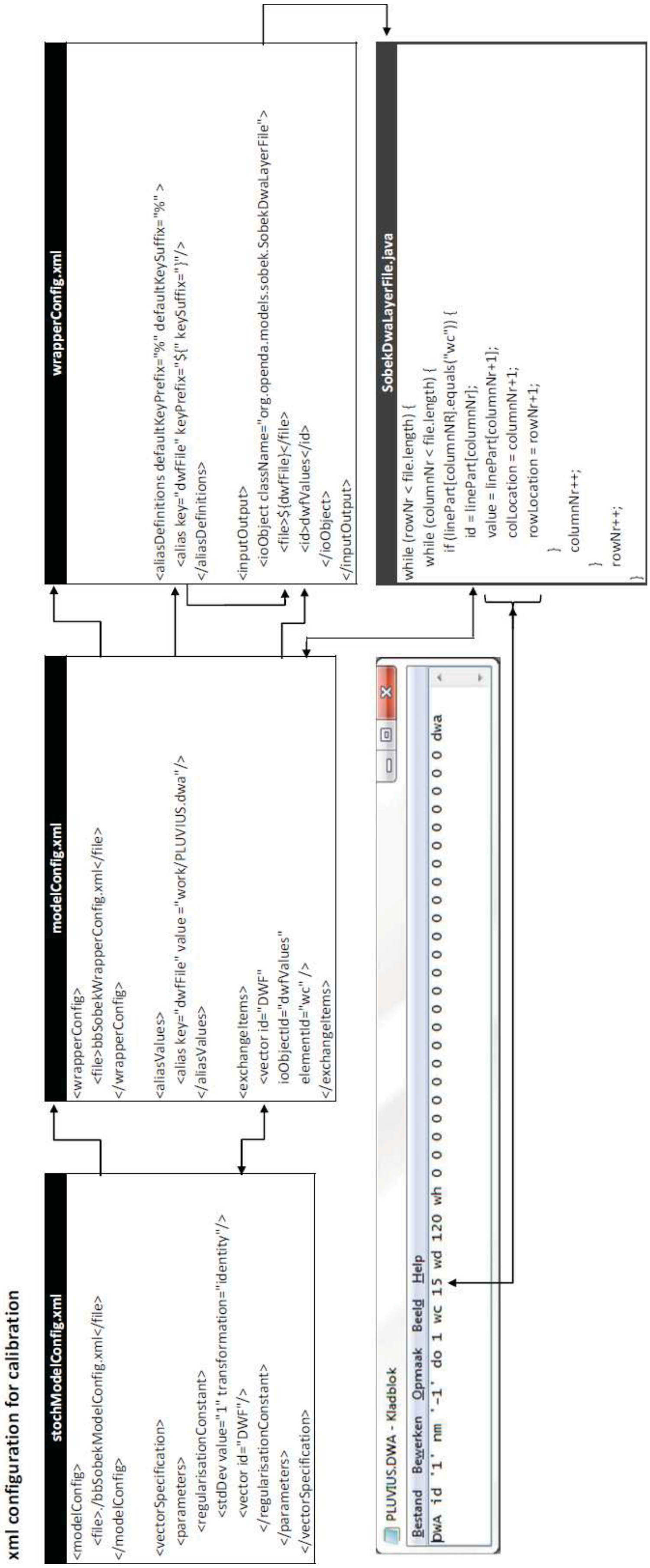


Figure 66: an example of Gaussian white noise added to observations produced by a model

Annexe VII: XML configuration



xml configuration for the EnKF

stochModelConfig.xml

```
<modelConfig>
  <file>./bbSobekModelConfig.xml</file>
</modelConfig>

<vectorSpecification>
  <state>
    <vector id="DWF"/>
    <noiseModel>
      <vector id="DWF"/>
      ..
    </noiseModel>
  </state>
</vectorSpecification>
```

modelConfig.xml

```
<wrapperConfig>
  <file>bbSobekWrapperConfig.xml</file>
</wrapperConfig>

<aliasValues>
  <alias key="dwfFile" value ="work/PLUVIUS.dwa"/>
</aliasValues>

<exchangeItems>
  <vector id="DWF"
  ioObjectId="dwfValues"
  elementId="wc" />
</exchangeItems>
```

wrapperConfig.xml

```
<aliasDefinitions defaultKeyPrefix="%" defaultKeySuffix="%" >
  <alias key="dwfFile" keyPrefix="${" keySuffix="}"/>
</aliasDefinitions>

<inputOutput>
  <ioObject className="org.openda.models.sobek.SobekDwaLayerFile">
    <file>${dwfFile}</file>
    <id>dwfValues</id>
  </ioObject>
</inputOutput>
```

PLUVIUS.DWA - Kladblok

Bestand Bewerken Opmaak Beeld Help

```
DWA id '1' nm '-1' do 1 wc 15 wd 120 wh 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 dwa
```

SobekDwaLayerFile.java

```
while (rowNr < file.length) {
  while (columnNr < file.length) {
    if (linePart[columnNR].equals("wc")) {
      id = linePart[columnNr];
      value = linePart[columnNr+1];
      colLocation = columnNr+1;
      rowLocation = rowNr+1;
    }
    columnNr++;
  }
  rowNr++;
}
```

Delft University of Technology
Faculty of Civil Engineering and Geosciences
Department of Water Management
Section of Sanitary Engineering
Stevinweg 1
2628 CN Delft

www.sanitaryengineering.tudelft.nl