OPTIMISATION OF MONITORING NETWORKS FOR WATER SYSTEMS

INFORMATION THEORY, VALUE OF INFORMATION AND PUBLIC PARTICIPATION

JOSÉ LEONARDO ALFONSO SEGURA
OPTIMISATION OF MONITORING NETWORKS FOR WATER SYSTEMS
INFORMATION THEORY, VALUE OF INFORMATION AND PUBLIC PARTICIPATION
OPTIMISATION OF MONITORING NETWORKS FOR WATER SYSTEMS
INFORMATION THEORY, VALUE OF INFORMATION AND
PUBLIC PARTICIPATION

DISSERTATION
Submitted in fulfilment of the requirements of
the Board for Doctorates of Delft University of Technology
and of the Academic Board of the UNESCO-IHE Institute for
Water Education for the Degree of DOCTOR
to be defended in public
on Tuesday 16th of November 2010 at 15:00 hours
in Delft, The Netherlands

by

José Leonardo ALFONSO SEGURA

born in Bogotá, Colombia.
Master of Science in Hydroinformatics
UNESCO-IHE Delft, the Netherlands
Dedicated with endless love

to my wife Sandra and our children

Valentina and Sebastian
Summary

A vital part of modern water management is the measurement of different processes of the hydrologic cycle. For this purpose, monitoring networks provide data that is analysed to help managers make informed decisions about a water system of interest. However, there are a number of main challenges regarding the design and evaluation of monitoring networks, which range from establishing proper temporal and spatial scales to defining their scope at minimum costs.

In theory, these difficulties are generally addressed by scientists when developing new approaches for monitoring network design. However, in practice, the data collected by the existing monitoring networks remains, in general, inadequate for understanding and explaining the dynamics of natural systems. Although this may be because the criteria to establish the final monitoring network is driven in practice by non-scientific aspects, such as political and social viewpoints, new approaches that involve the nature of the decision maker and stakeholder participation can reduce the gap between theory and practice.

In this Ph.D. research, funded by Delft Cluster, Delfland Water Board and UNESCO-IHE, innovative methods to design and evaluate monitoring networks are addressed. The main idea is to maximise the performance of water systems by optimising the information content that can be obtained from monitoring networks. When we talk about performance of water systems, we refer to the classification of how the water “behaves” with respect to each particular water use and interest within a particular system. The estimation of this performance should drive decisions made about the system. Additionally, when we talk about maximising information content, we refer to the task of seeking those potential locations where the placement of a set of data collection devices can give the best indication of the state of the system at any point within the water system.

Three stages form the pillars of the research: Information Theory, Value of Information, and data collection by public participation. The first pillar is an information theory-based method for localising water-level monitoring stations. The placement of a set of monitoring points that reduces the uncertainty, understood from the perspective of Shannon's information theory, of water-level measurements in a water system, has been achieved. Two main methods were developed. First is the Water Level Monitoring Design in Polders (WMP) method, in which the monitors are located one by one keeping the information content of the set as high as possible and the mutual information between pairs of monitors as low as possible. The second approach considers a Multi Objective Optimization Problem (MOOP), in which two practical situations are considered: 1) the costs of placing new monitors, and 2) the cost of placing monitors too close to hydraulic structures. The costs were considered in terms of informative units, for which additional terms affecting the objective functions were introduced. In addition, a third approach that considers the problem of locating discharge monitors in the Magdalena River by complementing the MOOP with a rank-based greedy algorithm, is presented.
The second pillar considers a method to place monitoring devices according to the value that the user of these devices gives to the information collected, according to its usefulness to make decisions. We use concepts of Value of Information, defined as the expected difference between the utility of choosing a particular action given the prior beliefs about the state of the system and the utility of choosing the action given additional information coming from an informative device (monitor). The approach selects the most valuable set of monitors for a particular water system, taking into consideration the decision-maker’s prior beliefs, the consequences of the system performance and the quality of the informational message that the monitoring network would provide. This new approach is characterised by a theoretical development for establishing the value of the location for one monitor and its extension to the case for multiple monitors. In addition, it is proposed that the probabilistic variables needed to calculate the Value of Information are estimated through the use of models.

The third pillar of the research, driven by a significant practical component, aims to explore new possibilities of gathering information with mobile phones and to improve models with this data. Nowadays mobile phones cannot be considered as merely telephones for voice transmission anymore: they have become devices that combine PC software, digital cameras, calculators and agendas and that are able to provide Internet, radio, television and fax services. In particular, mobile phones may be used by the public for water level monitoring. The idea is to take advantage of the benefits of public participation in monitoring, identified by diverse authors, which include the creation of public awareness on environmental issues, the improvement of collaboration among all stakeholders, the cost effectiveness of data collection activities and the high coverage in space and time. An important contribution in this respect is the Mobile Monitoring Experiment (MoMoX), an experiment carried out during 2010 in the polders of Pijnacker with participants of diverse affiliations, including residents near the water level gauges.

Two main case studies of very different nature are used to test the developed methods and theories. First, the polders of Pijnacker, a typical low-lying regional, flat and highly controlled system in the Netherlands, located to the East of the Delfland area. The region is mainly rural, with some urban development and greenhouses. From the hydrologic point of view, four major polders, subdivided into 127 smaller polders, are hydrologically independent response units, with unique target water levels. The system is controlled by a number of pumping stations, inlets and fixed weirs that are operated in order to keep the water levels in the canal network between limits defined by the water management of the region.

The second case study corresponds to the Magdalena River, the major river system in Colombia and the largest river discharging into the Caribbean Sea. It runs for about 1,530 kilometres from South to North, draining a catchment equivalent to 24% of the area of the country, and where 77% of the population lives. The area of interest for this research is the middle and low Magdalena, which are important for the country not only in terms of economic activities such as navigation, fish production and agriculture, but also because these regions suffer most from flooding.
The results of this research demonstrate that monitoring networks can be evaluated and designed by considering new criteria, such as the information content, the nature of the information user and the potential of the current mobile power for data collection. Among them, new methodologies for optimising monitoring networks coupling models with concepts of Information Theory and aspects of Value of Information have been achieved, and a public-based monitoring network for water-level data collection, characterised by the use of mobile phones has been configured, tested, run and assessed.

Additionally, this research opens new possibilities for the application of Information Theory in water resources, in which information has been traditionally quantified in entropy units. With the inclusion of Value of Information concepts, the entropy-based methods developed in this thesis, as well as the ones developed in previous studies, can now be adjusted to consider measuring the information in monetary terms.
## Contents

**SUMMARY** ............................................................................................................................................... VII

### CHAPTER 1  INTRODUCTION ........................................................................................................................... 1

1.1 BACKGROUND ........................................................................................................................................... 1
1.2 MOTIVATION OF THIS THESIS .................................................................................................................. 5
1.3 RESEARCH QUESTIONS AND OBJECTIVES ............................................................................................... 7
1.4 SCOPE OF THIS THESIS ........................................................................................................................... 7
1.5 THESIS OUTLINE ...................................................................................................................................... 9

### CHAPTER 2  LITERATURE REVIEW .............................................................................................................. 11

2.1 INTRODUCTION ........................................................................................................................................ 11
2.2 DESIGN AND EVALUATION OF MONITORING NETWORKS .................................................................... 11
2.3 INFORMATION THEORY (IT) .................................................................................................................... 13
2.4 VALUE OF INFORMATION (VOI) ............................................................................................................ 20
2.5 PUBLIC PARTICIPATION IN MONITORING ............................................................................................ 24
2.6 MODEL RELIABILITY ............................................................................................................................... 25

### CHAPTER 3  CASE STUDY 1: POLDERS OF PIJNACKER .............................................................................. 27

3.1 INTRODUCTION ....................................................................................................................................... 27
3.2 DESCRIPTION OF THE POLDER SYSTEM OF THE PIJNACKER REGION .............................................. 30
3.3 WATER LEVEL MONITORING NETWORK .............................................................................................. 33
3.4 DESCRIPTION OF THE MODEL OF THE PIJNACKER POLDER SYSTEM ............................................... 34

### CHAPTER 4  CASE STUDY 2: MAGDALENA RIVER ...................................................................................... 37

4.1 INTRODUCTION ....................................................................................................................................... 37
4.2 PERFORMANCE OF THE MAGDALENA RIVER ..................................................................................... 44
4.3 DEVELOPMENT OF THE HYDRO_DYNAMIC MODEL FOR THE MAGDALENA RIVER ...................... 45
4.4 LIMITATIONS OF THE MODEL ............................................................................................................... 55

### CHAPTER 5  INFORMATION THEORY FOR MONITOR LOCATION ............................................................. 57

5.1 INTRODUCTION ....................................................................................................................................... 57
5.2 INFORMATION THEORY-BASED APPROACH FOR LOCATION OF MONITORING WATER LEVEL GAUGES IN POLDERS ........................................................................................................... 59
5.3 OPTIMIZING INFORMATION MEASURES FOR THE DESIGN AND EVALUATION OF MONITORING NETWORKS IN POLDERS ............................................................................................................. 72
5.4 EVALUATION OF THE MONITORING NETWORK OF THE MAGDALENA RIVER .............................. 87
5.5 CONCLUSIONS ....................................................................................................................................... 100

### CHAPTER 6  VALUE OF INFORMATION FOR MONITOR LOCATION ......................................................... 103

6.1 INTRODUCTION ....................................................................................................................................... 103
6.2 DEFINITION OF VARIABLES FOR VOI ESTIMATION ............................................................................ 104
6.3 VALUE OF THE LOCATION FOR ONE MONITOR .................................................................................. 108
6.4 VALUE OF THE LOCATIONS FOR TWO MONITORS ............................................................................. 110
Chapter 1
Introduction

Everyday we measure things. We look at our watches in order to keep track of time to avoid running too late; we look at the petrol level to know whether we can reach our destination with the available combustible; we measure our weight to check how much time should we spend in the gym; we look into our pockets to know whether we have enough money to buy something (or to know whether it is time to ask for a salary raise).

In general, we measure because we want to make informed decisions about something. The path between measuring and making a decision is, however, not straightforward. Once we measure, we must interpret and analyse the data, and account for errors it may have. Then, we gain the information that we need to understand and to decide.

A vital part of modern water management is the measurement of the different processes of the hydrologic cycle. For this purpose, monitoring networks are situated to generate data with useful information content, which is used by managers of water systems and by decision-makers to maintain a good performance of their water system.

This chapter addresses the subject of this thesis and the importance of the research problem, and introduces the motivation to solve it.

1.1 Background

This section describes water systems and their performance, the importance of monitoring networks and their role in the modelling process, and introduces alternative means of data collection.

1.1.1 Water systems and their performance

Water systems
A water system can be defined as a set of interconnected components in which water is present and has demands made of it by diverse users with different interests. Examples of water systems are the natural rivers, coasts, lakes, and, in general, all the surface water and groundwater bodies, that are subject to interference by humans (to be used for example for agriculture, drinking water, hydropower generation, navigation, fishing,
recreation and industry) and required by nature in general (for instance terrestrial and aquatic ecology). In this thesis, approaches to designing, evaluating and optimising monitoring networks are applied to two water systems that are very different in nature: namely, rivers and polder systems.

Perhaps the most important water systems are the natural rivers, because of their scope, the number of users and interests involved from their source to their discharge to lakes or oceans and because of the complex physical and chemical processes that take place in their catchment. In fact, ancient civilizations settled in their floodplains and nowadays they are the focus of development and face enormous sustainability challenges, especially in developing countries.

In contrast, a completely different type of water system in terms of configuration (although not in terms of users and interests) is a polder system. A polder is a low-lying area that is artificially disconnected from the hydrological regime of neighbouring regions by means of hydraulic structures and dikes, in order to keep the water level within convenient ranges in the area of interest. Polders drain excess water independently into higher areas or to the sea.

Several differences between rivers and polders can be distinguished from different viewpoints. For example, in terms of area, what is drained by a polder system is considerably smaller than the area of a river catchment, which makes a difference also to the hydrological regimes in both systems; in terms of drainage type, rivers drain from higher to lower elevations, whereas polders require pumping stations to drain excess water to higher elevations; in terms of morphology, rivers “draw” their own stream path through natural processes, which may induce strong geomorphology changes in time, while the polder’s canal networks are built artificially, with low morphological changes.

Performance of water systems
When we talk about performance of water systems, we refer to the classification of how the water “behaves” with respect to each particular water use and interest. The estimation of this performance should drive decisions made about the system. According to Loucks et al. (2005) the performance of water systems can be quantified by selecting broad, general water management objectives and then splitting them into more detailed ones. If an objective is not general enough, then there is some risk of overlooking sub-objectives. As a general rule, broad objectives have the characteristic of being subject to a phrase like “because it is what we all want”. For example “protecting public health” and “increasing economic development” are objectives that everybody agrees with.

A major drawback, however, is that normally these water management objectives are conflicting. For example, water quality, the base for preserving the ecosystem and human health, is an issue that is frequently in conflict with human activities such as development and industrialization, which are the base of the regional economy; similarly, water quantity has different conflicts with different uses such as navigation (which requires high water levels) and agriculture (which requires specific water level ranges). This is the reason why it is so important to optimise the performance of water systems.
For the case of the Pijnacker region, these conflicts arise when managing water quantity and water quality. For instance, navigation, ecology, and recreation require high water levels, while flood prevention, horticulture and pasture agriculture prefer low water levels. Similarly, for the case of the Magdalena River, high water levels imply not only conflicts between flood management and cattle farming, but also between fishing production and illegal land-reclamation activities.

The idea behind measuring the performance of water systems is that decision-makers can make quantitative-based decisions, even if these decisions are qualitative.

1.1.1 Data, Information, Knowledge, Models

Data is not useful by itself. From the Knowledge Management viewpoint, when data is processed to make it useful we call it information, and when information is applied we call it knowledge (Ackoff 1989). Additionally, Abbott (2002) states that models are encapsulated knowledge that we can use to produce information.

However, models not only transform data into information (e.g., transforming channel characteristics and flow boundary conditions into stage and discharge at every computational point), but also need data to improve themselves (e.g., making sure the model replicates the measurements well through calibration and validation). Cunge (2003) offers an interesting discussion on why the relationship between data and models is much more complicated than what it is often presented by teachers and practitioners in the water sector. This discussion includes the role of data in calibration and validation of deterministic models, and proposes good practice for modelling.

Nowadays models are considered essential for decision making, since they provide a way for planners and managers to predict the behaviour of any planned water system design or management policy before it is implemented. They are also used to make decisions to optimally control the current water system, by operating control devices that keep water variables of interest within predefined ranges. In both cases, the water management objectives, with all their conflicts, should be taken into account (Lobbrecht 1997). Models can be used, therefore, to improve the performance of water systems.

However, before using them for any purpose, models must reliably replicate the state of the system under consideration and for this data is needed, so that calibration and validation can be performed. This condition implies that the data collection is a critical issue, since it is the first input within the decision-making process (data to generate information and information to generate decisions). For this reason, monitoring networks are of paramount importance.

1.1.2 Monitoring networks

A monitoring network can be defined as a set of strategically located measurement devices that collect data of interest about a water system at a given temporal scale. Monitoring networks are important because they collect data that, after being interpreted, provide insights for decision-making. In this thesis we understand monitoring as the process of observing what is happening in the water system.
Hydrological monitoring networks are, in a wider concept, part of environmental networks, which have been classified by Estrin et al. (2003), according to the type of variable measured, whether physical, chemical or biological. Regarding the function, the hydrological monitoring networks are conformed by devices that monitor the hydrosphere and the atmosphere by measuring the presence of water in the ground, on the surface and in the atmosphere. As a main objective of monitoring, Hooper et al. (2004) highlighted the understanding of four basic properties, not only referring to water, but also to sediments, nutrients and pollutants: mass, residence time, fluxes and flow paths. In this thesis we concentrate on monitoring networks for surface water levels and flows.

There is a close relationship between what is needed to manage and what is needed to be collected. General objectives of hydrological monitoring are completely linked with the objectives for the performance of water systems described above. The World Meteorological Organisation (WMO) stated in 1981 that “the aim of a network is to provide a density and distribution of stations in a region such that, by interpolation between data sets at different stations, it will be possible to determine with sufficient accuracy for practical purposes, the characteristics of the basic hydrological and meteorological elements anywhere in the region” (Made 1988, page 20). However, the use of models as interpolators is not explicitly mentioned in WMO guidelines, perhaps because of the existing difficulties for meteorological modelling.

The questions to be addressed when designing monitoring networks are what to measure, why, where, how often and with what accuracy (Loucks et al. 2005). First of all, a clear definition of the objectives of monitoring and the data needs is necessary. This definition of objectives leads to the monitoring plan, where it is explicitly defined which data is required, its accuracy and its frequency of monitoring. Temporal and spatial scales are defined in this plan as well. There exist many methods for design and evaluation of monitoring networks. Mishra and Coulibaly (2009) provide a comprehensive review of the available methods, namely statistical methods, entropy-based methods, basin physiographic characteristics and sampling strategies. In this review, it is demonstrated that statistical methods are the most developed.

1.1.3 Information Theory and Value of Information

The main concepts of Information theory were developed in the late 1940’s in the field of communication systems by Shannon (1948), with the aim of providing a quantitative measure of information. In brief, it states that any entity that provides surprising outcomes has more information than the entities that do not. For instance, telling the readers that the author speaks Spanish will not surprise them, as they might infer so just by looking at his name (or by hearing his English accent). Now, telling the readers that the author is a very skilled football player (even if this is not true) will be more surprising for the readers. Therefore, the statement “the author is a very good football player” adds more information to the reader than the statement “the author speaks Spanish”. One point to clarify in this respect is that Information Theory does not account for the meaning because “these semantic aspects of communication are irrelevant to the engineering problem”, (Shannon 1948).
In contrast, the main concepts of Value of Information were developed in the field of economics during the late 1960’s (Howard 1968) as it was of interest to know the amount a decision maker would be willing to pay for information prior to making a decision. An important characteristic of this theory is that the decision-maker’s beliefs are significant when assigning a value to information and that its value rises as the cost of making the wrong decision increases. For instance, consider a man who has to decide whether or not to go to the doctor when he feels unwell. If the subject is either a hypochondriac (he has an excessive worry about having a serious illness) or iatrophobic (he fears to go to the doctor), the decision is already made (a hypochondriac will certainly go and the iatrophobic will certainly not go). In both cases, any possible additional information to make the decision has no value, because the prior beliefs are at extremes. Conversely, if the man is completely doubtful about what to do, any additional information (maybe googling his symptoms) will be valuable for him to make a decision. Additionally, it is clear that wrongly deciding to go to the doctor (“paracetamol-water-call in two weeks”), has relatively less important consequences than wrongly deciding not to do it (further complications, life at risk).

1.1.4 Public participation in monitoring

Public participation is an interesting option for data collection. The benefits of public participation in monitoring have been identified by diverse authors for the case of water quality monitoring (see e.g., Au et al. 2000; Bromenshenk and Preston 1986; Stokes et al. 1990), and include the creation of public awareness on environmental issues, the improvement of collaboration among all stakeholders, the cost effectiveness of data collection activities and the high coverage in space and time.

Similarly, disadvantages of voluntary collection have also been identified and include the lack of confidence in data collection procedures, data quality, which is often unknown and data that is usually dispersed and non-structured. Gouveia et al. (2004) suggest the use of Information and Communication Technologies to overcome these problems and explore non-traditional types of environmental data such as images, sounds and videos. In this thesis, we explore the use of mobile phones for data collection by the public.

1.2 Motivation of this thesis

There are a number of main challenges regarding the design and evaluation of monitoring networks, which range from establishing proper temporal and spatial scales to defining their scope at minimum costs. In theory, these difficulties are generally taken into account by scientists when developing new approaches for monitoring network design. However, in practice, it has been reported that the data collected by the existing monitoring networks remains, in general, inadequate for understanding and explaining the dynamics of natural systems (Canadian Water Resources et al. 1994; IUCN 1980). This may be because the criteria to establish the final monitoring network is driven in practice by non-scientific aspects, such as political and social viewpoints.
Additionally, Mishra and Coulibaly (2009, page 19), point out that “...it is anticipated that the future will witness a greater and growing demand of hydrometric information for water resources, environmental, and ecohydrological management”. Provided the objectives of a monitoring network are defined, it is clear that the ideal, uncertainty-free monitoring network is one that has an infinite number of monitors\(^1\) in the area of interest, each one providing data at infinitesimal temporal scale. Naturally, this implies also the need for infinite resources for operation, management and analysis. Yet another challenge, then, is to properly select key locations where to collect the data needed to make proper decisions about the water system with the minimum expense. The cartoon in Figure 1-1 expresses the concern of not being able to reliably acquire the state of the system throughout the existing monitoring network.

\[\text{Figure 1-1. One of the main challenges when designing a monitoring network}\]

It is clear, then, that the resolution of these main challenges will imply on the one hand the generation of improved water management decisions with regards to the performance of water systems and on the other hand the possibility of enhancing our knowledge about the variability of natural systems.

\(^1\) In this thesis the terms monitor, device or gauge are used interchangeably to refer to any piece of equipment used to measure any water-related variable.
1.3 Research questions and objectives

In the context presented above, a main question arises:

*How can monitoring networks be optimised (in such a way that information content and value is maximized), and how can a public-based monitoring network be configured, with the aim of enhancing the performance of a water system?*

In order to give an answer to this question, the following research questions have been posed:

- How can models and Information Theory concepts be coupled to optimise monitoring networks?
- How can models and Value of Information concepts be coupled to optimise monitoring networks?
- How can a public-based monitoring network for hydrological data collection be configured and how can this data be used to reduce model errors so that an improved description of the water system, especially during extreme events, can be obtained?
- How can the new approaches be applied to polder systems (such as the Delfland system in The Netherlands) and to large river systems (such as the Magdalena River)?

The objective of this research is to investigate new methods of optimising monitoring networks, using concepts of Information Theory, Value of Information and public participation.

The particular objectives of this research are listed as follows:

- To explore different methodologies for optimising monitoring networks using Information Theory concepts and models.
- To develop a methodology for optimising monitoring networks using Value of Information concepts and models.
- To configure, test, run and assess a public-based monitoring network for water-level data collection, characterised by the use of mobile phones.
- To develop a methodology to use the mobile phone data collected by the public to reduce model errors.
- To test the developed methodologies and tools in two case studies in water systems with different hydrologic, hydraulic, socioeconomic and political conditions.

1.4 Scope of this thesis

1.4.1 Role of models in monitoring network design

There exist different ways to deduce data of non-measured sites as from the measured ones. Made (1988) provides examples of different interpolation methods for the case of designing and evaluating monitoring networks, which include mathematical relations, statistical methods, physically-based mathematical models and their combinations. In this
thesis, we use physically-based mathematical models to generate dense time series data sets from a limited set of data in order to account for the physics of the phenomena, allowing for reliable descriptions of the water system under different conditions.

Subsequently, the resulting data sets are used to design or evaluate the monitoring network, in such a way that the best points where measurements can be taken are identified. After this, the designed monitoring network, now in place, will generate data with a maximized information content (Chapter 5) and with a maximum Value of Information (Chapter 6), taking into account the predefined water management objectives for decision making. This data, in turn, can be used to improve the model through improved calibration and validation. The loop is closed when the improved model is again used to design or evaluate the previous monitoring network. The process is repeated until the monitor locations do not change between two consecutive loops. This iterative process is presented in Figure 1-2.

![Figure 1-2. Use of models for the design of monitoring networks (this thesis)](image)

### 1.4.2 Other means of data collection

Nowadays mobile phones cannot be considered as merely telephones for voice transmission anymore. They have become a device that combines PC software, digital cameras, calculators and agendas and that is able to provide Internet, radio, television and fax services. One of the most important facts regarding these devices is that any person may own one and that the knowledge of operating one is in the public domain. All of these characteristics make the mobile phone to be a cheap device not only for receiving data but also for sending all kinds of data at conveniently dense spatial and temporal scales. The option to collect data through mobile phones by means of public participation is explored in this thesis (Chapter 7). These new approaches for data collection have a potentially positive impact in developing countries, where mobile phone technology is
accessible to the poor people. A massive participation for data collection could lead to better water management of water systems that lack proper monitoring technology. The improved water management, in turn, will provide better living conditions to citizens, confirming the societal component of Hydroinformatics (Abbott 1991; Jonoski 2002).

1.4.3 Scope

The scope in which the questions posed before are answered includes the first loop of the flowchart shown in Figure 1-2 for the monitoring of water level and discharges. The monitoring networks are optimised spatially with theoretical approaches (Information Theory, Value of Information) and complemented temporally with practical methods (mobile phone-based public participation for data collection), the latter being also used to explore the possibilities for identifying the sources of error in a model.

1.5 Thesis outline

The research questions posed in the previous section are answered through the development of six interconnected chapters, which begin with Chapter 2, where the state of the art is provided through a detailed review of the literature and where the theories used in the course of the thesis are presented.

In order to develop the proposed methods, two case studies in catchments with different hydrologic, hydraulic, socioeconomic and political conditions, namely the Polders of Pijnacker, The Netherlands and the Magdalena River, Colombia, are presented in Chapter 3 and Chapter 4 respectively. In the latter, details about the development of a hydrodynamic model, a major task elaborated during this thesis, is also included.

Subsequently, the proposed new methods for designing, evaluating and optimising monitoring networks are presented and applied in each case study, for which the use of hydrodynamic models is significant. The concepts of Information Theory and Value of Information are applied in Chapter 5 and Chapter 6, respectively.

In order to explore new methods for data collection throughout the participation of the inhabitants residing in a particular water system, Chapter 7 is presented. The popular short message service (SMS) mobile phones can send, is exploited for both data collection and model improvement. An important contribution in this respect is the Mobile Monitoring Experiment (MoMoX), an experiment carried out during 2010 in the polders of Pijnacker.

Finally, Chapter 8 presents the conclusions and recommendations for each proposed method.

In order to help the reader to understand the relationship between the chapters, a schema of the thesis outline is provided in Figure 1-3. First, the Introduction and the Literature Review work as the base of the three main pillars of the thesis (Information Theory, Value of Information and Public Participation - Model Reliability). The case studies
Optimisation of monitoring networks for water systems

(Polders of Pijnacker and the Magdalena River) cross horizontally these pillars, showing where the developed methods are applied. Finally, the Conclusions are supported by the whole framework.

Figure 1-3. Outline of the thesis
Chapter 2
Literature Review

2.1 Introduction

Different issues regarding the monitoring of water systems are addressed in this thesis. For this reason, it is necessary to provide the theoretical framework for the methods developed further through a detailed review of the relevant literature. In the first place, the studies related to the design and evaluation of monitoring networks are presented, which include a number of existing approaches for the design and evaluation of networks of diverse nature. Subsequently, the basic expressions of Information Theory are presented along with a number of publications that apply the theory for water-related problems, including the design and evaluation of monitoring networks for different purposes. In particular, the review in Information Theory is necessary to introduce the methods developed in Chapter 5. Thirdly, a review of the concepts, mathematical development and the application of the Value of Information (VOI) concept is presented also in the water resources context, which is the starting point of the development of the methodology presented in Chapter 6. The starting point of the developments given in Chapter 7 is the combination of two important topics, namely the public participation in monitoring using Information and Communication Technologies (ICT) and model reliability and uncertainty. The relevant bibliography of both topics is reviewed at the end of this chapter.

2.2 Design and evaluation of monitoring networks

One of the most important elements of the planning and management of water resources is the assessment of these resources, which takes into account the identification of the sources and the evaluation of their capacity, dependability and quality, implying the measurement and collection of data of interest (Mishra and Coulibaly 2009). For this purpose, monitoring networks are designed and sometimes optimized for decision-making according to the water management objectives (Loucks et al. 2005). In the following sections recent developments in the design and evaluation of monitoring networks are discussed.

2.2.1 Designing a monitoring network

As is the case for any water system, the estimation of water-related variables at sites where no measurements take place is the main problem the design of a monitoring
network must address. Therefore, the design of a monitoring network aims to find the number and spatial distribution of gauges and the temporal interval of their measurements.

The monitoring networks must be planned in a process that is not static, but that requires a continuous cycle. Figure 2-1 shows a simple cycle that Mogheir and Singh (2002) indicated as a first approach monitoring cycle. In summary, information needs come from what has been defined as important to manage in the water system. The information strategy is, thus, the subsequent step that leads to a plan for data collection. After this, the analysis of data and the utilisation of the information are the processes that create the inputs for water management and decision making, closing the cycle in this way.

A detailed description of the process for designing monitoring networks is given by WMO (1994), which suggests beginning with the definition of the institutional setup and the purposes, objectives and priorities of the network. Subsequently, the obtained network design is under optimization procedures, financial revisions and final implementation. The last step includes a review of the network which, together with other feedback mechanisms applied at each of the previous steps, complete the framework.

![Figure 2-1. First cyclic approach for monitoring planning](image)

*Adopted from Mogheir and Singh (2002)*

A number of approaches to monitoring network design can be found in Mogheir et al. (2006), who classified them as geostatistical and statistical methods. Among these, variance-based, probability-based and entropy-based methods can be mentioned. A comprehensive review is presented in the work by Mishra and Coulibaly (2009).

Although there is extensive literature on a number of approaches to rain-gauge network design (see e.g., Bogardi and Bardossy 1985; Bras and Rodríguez-Iturbe 1976; Moore et al. 2000; Pardo-Igúzquiza 1998; Rodríguez-Iturbe and Mejia 1974; Sansó and Müller 1997; Yeh et al. 2006), there is comparatively little for the design of water level gauge networks. Some examples such as interpolation methods to find the minimal spatial density (Gandin 1965) and the number of stations required for runoff estimation (Karasev 1968), exist. Moss (1974) present an approach for the design of surface water data networks that include the statistical nature of parameter accuracy estimates, but this was presented as an application of a more general approach described by Moss (1976) for
hydrological monitoring networks. Similarly, Moss and Tasker (1991) performed a
collection between techniques for designing monitoring networks of stream-gauges, in
which the main variable was the discharge as an estimation of water levels. Husain
(1989) proposed a methodology to select the most important monitoring station out of a
dense set of stations, and also to expand hydrologic networks that have sparse stations. In
particular, he used gamma distributions to estimate the multivariate probability functions
in order to calculate the entropy-based information transmission.

From the design and operation point of view, Made (1988) analyzed a number of methods
to derive water levels at any point along river reaches, provided water level
measurements are available at existing points.

2.2.2 Evaluation of monitoring networks

Existing monitoring networks are evaluated in order to confirm that the objectives for
which the network was designed are met. The result of this assessment may include a
redefinition of the size and scope of the network, which can lead either to the elimination
(due to redundancy or uselessness of the collected data) or the inclusion of additional
monitoring points in places where water-related variables cannot be adequately inferred
from the existing monitors.

In general terms, the same approaches used for the design of monitoring networks are
used for their evaluation. These approaches can be classified as statistically based
methods, information theory methods, user survey approach, hybrid methods,
physiographic components and sampling strategies, which are presented in detail by
Mishra and Coulibaly (2009) and schematized in Figure 2-2.

Two approaches for the design and evaluation of monitoring networks developed in this
thesis are part of the entropy-based methods (see Chapter 5), two of them using
Optimization methods. Yet an additional approach that is not included in the review of
Mishra and Coulibaly (2009), is one that takes into account the Value of Information and
that is also developed in this thesis (see Chapter 6).

2.3 Information Theory (IT)

From the beginning of the development of Information Theory, two problems regarding
information concepts frequently came into view: what is information and how can it be
measured? Even though the first question seems to be crucial for any theory, the second
has been the most developed (Burgin 2003).

An interesting, general definition of information is given by Yankovsky (2000): “Any
interaction between objects during which one object gains some substance and the other
does not lose it, is called Information Interaction. In this case, the substance under
transmission is referred to as Information.”
Generally speaking, all developed theories to measure information can be considered as sub theories of the General Theory of Information (Burgin 2003). Among these theories, Shannon’s information theory, semantic theory of information, Fisher Information, qualitative information theory, algorithmic theory of information, pragmatic theory of information, social information, utility theory of information, economic theory of information and dynamic theory of information can be found.

The self-information introduced in the Theory of Information (Shannon 1948), the Kullback-Leibler divergence (as a measure of the differences between a ‘true’ and an ‘arbitrary’ probability distribution), and the Fisher Information (as the amount of information that an observable random variable carries about an unknown parameter upon which the likelihood function of the random variable depends) are related approaches for measuring information. In this thesis the Theory of Information is used to develop the methods presented in Chapter 5.
2.3.1 Quantifying information

Information theory as described by Shannon (1948), provides mechanisms for measuring information $R$, which is a reduction in uncertainty $H(X)$. As the latter is also known as entropy, information entropy, Shannon entropy or marginal entropy, the terms entropy and uncertainty will be used indistinctly through this thesis. The definition of uncertainty indicates how surprising is, in average, to get a value $x$ from a random variable $X$ that can take the possible values $x_1, x_2, ..., x_n$ each with probability $p(x)$ (Equation (2-1))

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$$  \hspace{1cm} (2-1)

The units of uncertainty are actually given by the base of the logarithm utilized, being “nats” if the base is $e$ and “bits” if it is 2. In this thesis the latter will be used. Another important consideration is that $0 \log 0 = 0$, which is in line with the fact that $x \log x \to 0$ as $x \to 0$ and values with zero probability do not change the uncertainty. An analysis of Equation (2-1) shows that when all values $x_i$ are equal, then the uncertainty of the variable is zero, thus the variable is not random and we are sure what the value $x_{eq}$ will be. On the other hand, when all possible values $x_i$ are equally likely then the uncertainty is maximum (Cover and Thomas 1991). This sense of the predictability of entropy-related expressions has been exploited in diverse fields such as climate (Majda et al. 2002), financial time series (Molgedey and Ebeling 2000) and DNA-sequences (Ebeling and Frommel 1998) among others.

Information $R$ is defined as a reduction in uncertainty Equation (2-2)

$$R = H_{before} - H_{after}$$  \hspace{1cm} (2-2)

In a flawless communication process (without noise), the receptor is completely certain about the message that was sent by the emitter, so $H_{after} = 0$ and $R = H_{before}$. This is the reason why some authors consider (wrongly) that information is the same as entropy or uncertainty (Schneider 2000).

It is possible to see that entropy is also an alternative measurement of variability or dispersion, just like the classical variance, but with some additional advantages: variance is not appropriate to use when the size of the sample is too small (Hart 1971; Mogheir et al. 2006; Singh 1997; Singh 2000), a situation that is handled by entropy; for discrete distributions without numerical values the variance becomes totally improper as the use of different numbered labels is translated into different values of variance, an inconvenience that entropy prevents, since only probabilities are taken into account. A more detailed discussion can be found in the work of Wei (1987).
In some cases, it is necessary to estimate the amount of information content between two random variables $X$ and $Y$. A frequently used approach is the mutual information (or transinformation) $I$, which quantifies the amount of information of one random variable which is contained in another random variable (Cover and Thomas 1991), and can be interpreted as the reduction of uncertainty of $X$ due to the knowledge of $Y$:

$$I(X;Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \quad (2-3)$$

Although the concept of transinformation and its relation with entropy can be depicted using Venn diagrams, they may not represent positive quantities when analyzing more than 2 variables (MacKay 2003)

Provided that $p(x, y)$ is the joint distribution between the variables $X$ and $Y$, and that $p(y \mid x)$ is the probability of $y$ given $x$, other information-related measures for two variables can be defined (Cover and Thomas 1991) including the joint entropy, given by:

$$H(X,Y) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log p(x_i, y_j) \quad (2-4)$$

which represents the amount of information that is contained in both variables, and the conditional entropy, given by:

$$H(Y \mid X) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log p(y_j \mid x_j) \quad (2-5)$$

which represents the amount of information content of $X$ which is not contained in $Y$.

In communication systems, $H(X)$ is the information input at the source $X$, $H(Y)$ is the information output at the receiver $Y$ and $I(X,Y)$ is the amount of information transferred from $X$ to $Y$ (MacKay 2003). It is interesting that $H(X \mid Y)$ is the amount of information loss during the transmission (part of $X$ that never reaches $Y$) and that $H(Y \mid X)$ is the amount of information that is received as noise (part received by $Y$ that was never sent by $X$). It is clear that neither of these values can be negative.

Although there are a number of approaches to measure the dependence between variables, the use of the mutual information has become popular in several fields of science for a number of reasons. First, because it does not take into account any relational hypothesis between the variables beforehand, as for instance, the linearity of the Pearson correlation (Steuer et al. 2002) or the linearity and normality of the ordinary correlation coefficient $r$ (Singh 2000) do. In other words, mutual information does not measure linear
dependencies but general dependencies, which are more likely to exist in physical and other natural processes. Second, as in the case of entropy, correlation functions can only be applied to a sequence of numbers, whereas mutual information can also be applied to a sequence of symbols (Li 1990). Furthermore, the invariance under transformation provided by entropy-based relationships is another advantage reported over the ordinary correlation (Linfoot 1957), as well as the capability to provide not only quantitative measure of information of one gauge station, but also a measure of the transference and loss of information between them (Yang and Burn 1994).

One major restriction of the mutual information is that it is only applicable to two random variables, and often it is needed to evaluate the dependency among several variables. This situation is addressed by the multivariate mutual information, a topic first studied by McGill (1954), who defined interaction information (or co-information), for the case of three variables, as:

$$I(X; Y; Z) = H(X) + H(Y) + H(Z) - [H(X, Y) + H(Y, Z) + H(X, Z)] + H(X, Y, Z)$$

$$= I(X, Y; Z) - I(X; Y) - I(Y, Z)$$

(2-6)

The general expression for $N$ variables, which was extended by Fano (1968) and reformulated by Han (1980), is:

$$I(X_1; X_2; \ldots; X_N) = I(X_1; X_2) - I(X_1; X_2; \ldots; X_{N-1} \mid X_N)$$

(2.7)

Srinivasa (2005) interprets interaction information as the gain (or loss) in the information transmitted between a set of variables due to additional knowledge of a new variable, whereas Jakulin and Bratko (2003) think that it is a measure of the amount of information that is common to all the variables, but not present in any of them. Fass (2006) interprets it as the influence of one variable on the amount of information shared between the rest of the variables. These interpretations are related to the fact that interaction information can be negative, because the dependency among a set of variables can increase or decrease with the knowledge of a new variable. For the case of knowing the effect of a third variable on the correlation of two variables, Jakulin and Bratko (2004) explains a positive interaction information as a synergy between the original variables and a negative value as a redundancy among these variables. Similarly, Fass (2006) states that after knowing the third variable, a resulting positive value “facilitates” or “enhances” the correlation between the two variables, whereas a negative value “inhibits” or “explains” this correlation. This author also recognizes that the difficulties in the interpretation of the possible negativity of the interaction information have been a barrier for its application in areas like machine learning and psychology (Fass 2006).

The concept of Total Correlation (McGill 1954; Watanabe 1960), $C(X_1, X_2, \ldots, X_N)$, provides a direct and effective way of assessing the dependency among multiple variables.
Optimisation of monitoring networks for water systems

\[ C(X_1, X_2, ..., X_N) = \sum_{i=1}^{N} H(X_i) - H(X_1, X_2, ..., X_N) \]  \hspace{1cm} (2-8)

It can be noted that for the case of \( N=2 \) that Total Correlation is equivalent to the well-known transinformation (or mutual information). The term \( H(X_1, X_2, ..., X_N) \) is the multivariate joint entropy (Eq. (2-4) for the case of \( N \) variables) of the set \( X_1, X_2, ..., X_N \). Total Correlation can be calculated by following the grouping property of mutual information (Kraskov et al. 2003), for the case of three variables \( X, Y \) and \( Z \), as is follows:

- A new variable \( A \) is built up by agglomerating \( X \) and \( Y \) in such a way that \( H(A) = H(X,Y) \). The procedure of ‘agglomeration’ consists of placing in \( A \) a unique value for every unique combination of the corresponding records in \( X \) and \( Y \). For instance, if \( X=[1,2,1,2,1] \) and \( Y=[2,3,1,3,2] \), then one of the options to agglomerate \( X \) and \( Y \) to build the variable \( A \) is by putting all the corresponding digits (or symbols) of \( X \) and \( Y \) together, i.e., \( A=[12,23,11,23,12] \).
- Following the same concept, a new variable \( B \) is built by the agglomeration of \( A \) and \( Z \), also with the condition \( H(B) = H(A,Z) \).
- The mutual information between the selected pairs for agglomeration, i.e., \( C(X,Y) = H(X) + H(Y) - H(A) \) and \( C(A,Z) = H(A) + H(Z) - H(B) \) are calculated.
- The Total Correlation of \( X, Y \) and \( Z \) is calculated by summing up the partial total correlations obtained for each built variable, i.e., \( C(X,Y,Z) = C(X,Y) + C(A,Z) \).

As can be noted, this method does not need to assess \( H(X_1, X_2, ..., X_N) \) and therefore the estimation of the joint probability distribution \( p(x_1, x_2, ..., x_N) \) is not needed. After having calculated \( C(X_1, X_2, ..., X_N) \) by following the steps above, the multivariate joint entropy \( H(X_1, X_2, ..., X_N) \) can then be calculated from Eq.(2-8) as:

\[ H(X_1, X_2, ..., X_N) = \sum_{i=1}^{N} H(X_i) - C(X_1, X_2, ..., X_N) \]  \hspace{1cm} (2-9)

A complete reference to related work can be found in the work of Jakulin and Bratko (2004) and Fass (2006). The application of the Total Correlation in water resources problems was first achieved by Alfonso et al.(2010b), a work presented in detail in section 5.3.

2.3.2 Uses of IT in the design and evaluation of monitoring networks

The main idea behind the design of any monitoring network is to reduce as much as possible the uncertainty associated with the estimation of values of a given variable in the places where it is not directly measured. The concept of uncertainty has been traditionally
linked with statistical variance, even though Amorocho and Espildora (1973) noticed that it was not an objective index of quality when used, for instance, for comparing predicted values of a hydrological model and the series of data records. It is interesting to note that during the same period, a similar observation was made in the field of portfolio management by Philippatos and Wilson (1972).

Diverse authors have applied information theory concepts to the design or evaluation of monitoring networks for general purposes (Caselton and Zidek 1984) and for more specific purposes, such as water quality (Harmanciovglu 1999), groundwater quality (Caselton and Husain 1980; Mogheir et al. 2004; Mogheir and Singh 2002; Mogheir et al. 2006), air pollution (Zidek et al. 2000) and rainfall gauging stations (Krastanovic and Singh 1992a; Krstanovic and Singh 1992b), among others.

From the point of view of surface water gauging, Husain (1989) presented a method for network design based on the information-transmitting capabilities of a hydrologic network in terms of entropy. Yang and Burn (1994) criticized this method, pointing out that a continuous distribution function is assumed when calculating the entropy-related measurements. Even though these authors overcame this problem by using a non-parametric estimation of the density distributions, the assumptions of having independent and identically distributed random variables and the assessment of the smoothing factor of the kernel parameter still continue to add vagueness to the process. These authors, nevertheless, presented a normalized version of the mutual information between two gauges, called the Directional Information Transfer Index (DIT), to obtain the fraction of information transferred from one site to another as a value between zero and one:

$$DIT_{X,Y} = \frac{I(X,Y)}{H(X)}$$  

(2.10)

The expression was first introduced in 1970 by Coombs, Dawes and Tversky in the field of Mathematical Psychology under the name Coefficient of Constraint (Fass 2006). Markus et al. (2003) explained the expression in Equation (2.10) as the information received by $X$ from $Y$. When $H(Y)$ is used in the denominator of Equation (2.10) they interpret it as the information sent from $X$ to $Y$. However, these interpretations are confusing since $DIT$ does not provide a quantification of the transmitted information content but a normalized version of the transinformation between these variables. For this reason, when the net information transfer $N$ is introduced as the difference between information sent and information received (Markus et al. 2003), negative values, that cannot be seen as an information gain whatsoever, are obtained.

Markus et al. (2003) presented a comparison between entropy and the least square method to evaluate stream gauges, in which the $DIT$ of Yang and Burn (1994) was adopted. The authors faced the problem of selecting the bin size for calculating the empirical frequency analysis to obtain the probability of a value in a particular interval. They found that, in spite of the differences in entropy values when changing the bin size, the ranking of stations in terms of the difference between the information received and
the information sent, remained, in general, inalterable. The problem of the bin size was pointed out from the beginning by Amorocho and Espildora (1973) and has also been studied in the case of the mutual information calculation of discrete variables by Steuer, Kurths et al. (2002).

Two difficulties appear recurrently in the studies that use entropy for monitoring network design. First, there is a problem of establishing the joint probability functions to calculate mutual information. This has been mainly solved assuming either a Gaussian distribution of the variables or evaluating the transinformation as a function of the correlation coefficient $r$, as suggested by Harmancioglu and Yevjevich (1987). Secondly, for the multivariate case, several simplifications are made, for example, analyzing mutual information by pairs of stations and analyzing the resulting 2D transinformation matrices (Filippini et al. 1994; Mogheir and Singh 2002), or assuming a normal distribution to calculate the multivariate joint entropy (Krstanovic and Singh 1992).

Estimating the joint probability of multiple variables is a problem encountered in several fields and diverse methods of approximation have been proposed. The problem is due to the combinatorial explosion of the number of probabilities to calculate for a large number of variables. Fass (2006), in his comprehensive literature review shows a number of proposed approximations, and states that this is a quintessential problem in human and machine learning. One of them is the Chow-Liu tree (Chow and Liu 1968), which approximates the joint probability as a product of bivariate distribution functions; the paired-variables with the biggest transinformation are selected to be part of the tree. Kirshner et al. (2004) applied this method to evaluate discrete time-series for modeling and forecasting daily precipitation occurrence for networks of rain stations, and demonstrated some improvements over simpler alternatives such as assuming conditional independence of the multivariate outputs.

Even though the method of total correlation expressed in Equation (2-8) for evaluating information among multiple variables is an indirect approach to evaluate multivariate joint probability functions, it is a direct, precise method to evaluate information dependency among multiple variables.

### 2.4 Value of Information (VOI)

The Value of Information is about how choices made under uncertain conditions affect the outcome. The basic idea is that a decision maker, conscious of his limited knowledge to make an informed decision (this is, under uncertainty), is willing to pay for additional information provided the expected gain exceeds the cost of collecting and analysing it. As stated by Weinberger (2001), “...the practical value of information derives from its usefulness in making informed decisions”.

The value of information depends on several factors (Macauley 2005):

- The degree of uncertainty of the decision-maker. If there are a few possible choices of remedial actions to decide from, then information has a small value even if it eliminates uncertainty. Conversely, if the costs of the actions have a
high variance (widely diverge), then the information can have a very high value even if it does not reduce uncertainty.

- The objects or issues that are at risk as an outcome of a decision; VOI depends on the value of the outcome. For instance, a willingness to pay for data about oil exploration potential is in part a function of the price of gas. Outcomes can also be measured in terms of damages caused by floods or diseases caused by pollution, when no goods or services are considered.
- The cost of using the information to make decisions; analysing (and not collecting) data can be so expensive that it is made to have little value.
- The price of the next-best alternative sources of information; sometimes there are several other substitutes for information (aerial photographs instead of satellite images, for example).

### 2.4.1 Definition

The concept of Value of Information was introduced first in the area of economy as a way to deal with the limitations of knowledge in the decision-making process. Hirshleifer and Riley (1979) offer a classic overview of general approaches to understanding the Value of Information. Following their theoretical development, the value of information, VOI, provided by one message received can be estimated as the difference between the utility, $u$, of the action, $a_m$, that is chosen given a particular message, $m$, and the utility of the action, $a_0$, that would have been chosen without additional information:

$$VOI_m = u(a_m, p) - u(a_0, p)$$  \hspace{1cm} (2.11)

Both utilities are calculated following the expected utility rule of Newmann-Morgenstern, by summing up the products of probabilities, $p$, and the consequences, $c$, resulting from the selected action, $a$, as shown in Eq. (2.12).

$$u(a, p_s) = \sum_{s} c_{as} p_s$$  \hspace{1cm} (2.12)

where, $p_s$, is the perceived probability of state, $s$, $c_{as}$ is the consequence associated with the decision of performing the action, $a$, when the system has a state, $s$, and $S$ is the set of possible states of the system.

On the one hand, if a decision maker has to make a decision without information, he/she will have to rely on his/her perspective about the state of the system. This perspective can be quantified as the probability $p_s = \pi_s$ of having a particular state, $s$, in the system. On the other hand, when new information is available, the decision maker can judge whether or not to believe the new information, thus updating his/her beliefs, or to reject the new information. Naturally, the new information has value if the individual uses it to make a decision, and has no value otherwise. Mathematically, the process for which the decision maker accepts the new information (i.e., is willing to change his/her belief, $\pi_s$) can be represented by Bayes theorem for updating belief as:
Optimisation of monitoring networks for water systems

\[ \pi_{s,m} = \sum_s \pi_s q_{m,s} \]  
\[ \text{(2.13)} \]

where \( \pi_{s,m} \) are the posterior or updated beliefs and \( q_{m,s} \) is the conditional probability of receiving the message, \( m \), given the state, \( s \), which is an amount that is generally provided by the information service, and is an indication of the quality of the message provided. False positives and true negatives, or error type I and error type II, are accounted for here.

In summary, Eq. (2.14) shows that the Value of Information is, ultimately, a function of \( c_{a,s} \), the consequences of taking an action, \( a \), given a particular state, \( s \); \( \pi_s \) is the prior probability, or the belief before the additional information; \( q_{m,s} \), the conditional probability of receiving the message, \( m \), given the state, \( s \).

\[ VOI = f\left(c_{a,s}, \pi_s, q_{m,s}\right) \]  
\[ \text{(2.14)} \]

Figure 2-3 presents the steps for the estimation of the Value of Information.

One of the difficulties in using the VOI lies in how to assess the probabilities before and after receiving the new information. Some authors have tried to estimate them empirically by interviewing decision makers directly (see e.g., Bouma et al. 2009; see e.g., Schimmelpfennig and Norton 2003), while other authors have used model outputs (see Dakins et al. 1996; Lin et al. 1999). The work by Yokota and Thompson (2004) offers a comprehensive review of various VOI applications in the field of environmental health risk management decisions.

### 2.4.2 Use of VOI in water resources

The concept of value of information has been applied in different fields. Some examples are the assessment of supply chains (Gavirneni et al. 1999; Lee et al. 2000), to explore bidders' incentives to gather information in auctions (e.g., Milgrom and Weber 1982), in industrial purchasing decisions (Stigler 1961), as a method for system identification (Fogel and Huang 1982), to assess the market distortions in real estate transactions (Levitt and Syverson 2008), among others.

The concept of VOI has been actively applied in the field of water quality monitoring. Some recent researches include the work by Ammar and Kaluarachchi (2009), who presented a methodology to optimise groundwater quality monitoring networks, taking into account vulnerability/probability assessment, environmental health risk, the value of information (VOI), and redundancy. Bouma et al. (2009) presented an assessment of the Value of Information for water quality in the North Sea, by combining Bayesian decision theory with an empirical, stakeholder-oriented approach.
Figure 2-3. Flowchart for VOI estimation
Shaqadan (2008) developed a framework to reduce the uncertainty in exposure to health risk due to drinking contaminated groundwater. The author assessed the socioeconomic value of potential decisions of collecting additional information for given variables, in which advanced social welfare concepts to understand the social acceptability of decisions to collect better information were presented.

In order to contain a plume of groundwater contamination through the installation and operation of pumping wells, a method by Wagner et al. (1992) was developed in which the hydraulic conductivity of the aquifer is the main source of uncertainty. Although these authors mention VOI as a way to evaluate the decisions, the concept is reduced to the solution of stochastic programming models and the evaluation of expected values of optimal solutions. Reichard and Evans (1989) examined the value of groundwater monitoring in reducing exposure uncertainty for different monitoring strategies.

Other uses of VOI in water resources include surface water quality, such as Borisova et al. (2005) who examine the price and quantity of different devices and assess the expected value of the information obtained for agricultural nitrogen pollution control; Ramirez et al. (1988) who introduce two analytical tools for decision making in flood control design, namely ex-post analysis and the Value of Information. Roberts et al. (2009) explores the value of information of early warning systems that detect contaminant agents of soybean crop and found that it mainly depends on the perceived risk of being infected and the accuracy of forecasts.

The concept of Value of Information is applied in this thesis to develop a methodology for placing monitors in a water system (see Chapter 6).

### 2.5 Public participation in monitoring

From the environmental standpoint, public participation is becoming an interesting option for the control and management of the environment. A number of examples of public participation in monitoring can be identified. Au et al. (2000) developed a methodology to involve the public in the collection of water quality data and found that high school students were able, after proper training, to reliably produce values for total coliforms, Escherichia coli and toxicity in waterways. In a similar project, Marcelino (2007) explored the possibilities of working with children to create a multisensory geographical information using Google Earth. This project was further extended to use mobile phones learning and participatory contexts (Silva et al. 2008).

Regarding the environmental monitoring, the framework Environmental Collaborative Monitoring Networks (Gouveia and Fonseca 2008) was created to channelize existing citizen initiatives that use ICT tools to increase the contribution of volunteered geographic data. Niinioja et al. (2004) described in detail the results of public participation in the monitoring of algae in the Lake Pyhäjärvi, on the border between Finland and Russia. Gouveia et al. (2004) suggested to overcome the problems of using voluntary collected (lack of confidence in data collection procedures, data quality often unknown and data usually dispersed and non-structured) by promoting the use of
Information and Communication Technologies with an emphasis given to tools that explore non-traditional types of environmental data such as images, sounds and videos in association with spatial information. Nare et al. (2006) assessed the extent of stakeholder participation in water quality monitoring and surveillance at the operational level, and also indigenous knowledge and practices in water quality monitoring in Zimbabwe, where policies and legislation encourage stakeholder participation.

In particular, this thesis concentrates on the use of mobile phones by the public for water level monitoring (see Chapter 7). The idea is to take advantage of the benefits of public participation in monitoring, identified by diverse authors for the case of water quality monitoring (see e.g., Au et al. 2000; Bromenshenk and Preston 1986; Stokes et al. 1990), which include the creation of public awareness on environmental issues, the improvement of collaboration among all the stakeholders, the cost effectiveness of data collection activities and the high coverage in space and time.

2.6 Model reliability

Models are a simplification of the reality and this simplification implies that models are not perfect. It is of particular interest, therefore, to assess the reliability of the models that are being used for planning, design or operation of water systems. In this regard, it is important to look at the sources of model uncertainty and review some of the available methods for uncertainty assessment.

2.6.1 Model uncertainty

Shrestha (2009) summarized the most important sources of uncertainty in rainfall runoff models, namely observational uncertainty, model uncertainty and parameter uncertainty. The first refers to the uncertainty of the measurements used by the model inputs and outputs; the second is associated to the simplifications, approximations and/or the lack of knowledge when building the conceptual structure to describe the physical processes of interest; the third is related to the model parameters that are generally estimated by indirect means such as expert judgment or calibration.

Following the definition given by Holling (1978), reliable models provide the confidence to base management decisions. It can be then stated that models are reliable when the outputs and their associated uncertainty are within acceptable limits. Therefore, one way to enhance the reliability of a model is to work on the reduction of its uncertainty.

2.6.2 Methods to assess uncertainty

There exist a significant number of studies to assess the uncertainty of rainfall-runoff models. A comprehensive review of the available methods is presented by Pappenberger (2006), where a decision tree is provided to select the proper method for a given situation. Five sets of methods can be mentioned: forward uncertainty propagation, model calibration and conditioning uncertainty, qualitative methods, real-time data assimilation and sensitivity analysis. The first method includes error propagation equations (see e.g., Kunstmann and Kastens 2006), Monte Carlo propagation (EPA 1997), reliability methods
(e.g., Melchers 1999) and fuzzy (e.g., Klar and Smith 2001) and imprecise methods (Walley 1991); the second includes nonlinear regression (e.g., Kavetski et al. 2002; Kuczera and Parent 1998), Bayesian methods (e.g., Van Oijen et al. 2005), and Generalized Likelihood Uncertainty Estimation, GLUE (e.g., Aronica et al. 2002); the third set of methods include the Numerical, Unit, Spread, Assessment and Pedigree, NUSAP (Jeroen et al. 2005) methods.

For real-time data assimilation, the Kalman filter-related methods (e.g., Bertino et al. 2003; Moradkhani et al. 2005b; Vrugt et al. 2005) and the sequential Monte Carlo analysis can be mentioned (e.g., Moradkhani et al. 2005a); regarding sensitivity analysis the works by Hall et al. (2005) and Wagener et al. (2003) can be cited.

In this thesis, a method to use data collected from the public to improve model reliability is presented in Chapter 7.
Chapter 3
Case study 1: Polders of Pijnacker

Two water systems of a completely different nature have been selected to develop the methods for monitoring network design, namely a polder system, in order to take into account a flat, highly controlled water system and a major natural river in order to consider uncontrolled, stream flows. For the first case, the polder system of the region of Pijnacker, The Netherlands, is described in this chapter. For the second case, the Magdalena River, Colombia, is introduced and described in Chapter 4. The methods for locating monitors are presented in Chapters 5, 6 and 7.

This chapter begins with a description of the general characteristics of a polder system and the particular strategy of the Delfland Water Board, the local authority that manages the polders of Pijnacker. This is followed by a description of the water system itself, including the characteristics of the drainage, the canal network and the control structures. Next, a description is given of the existing monitoring network, followed by a summary of a model of the water system, including a review of the rainfall-runoff and hydrodynamic models.

3.1 Introduction

Polders are developed areas below sea level, which are drained by canals and pumping stations that discharge excess water either to elevated storage basins or to the sea. In this way, artificial catchments are formed consisting of the combination of a number of polders normally delimited by weirs, pump stations or inlet structures.

The Netherlands, with about 20% of its territory below sea level, is famous for its polder water system configuration. A typical example is the western part of the country, in the province of Zuid Holland, where important cities such as The Hague and Rotterdam are located (Figure 3-1).

Local government agencies called "waterschappen" (water boards) or "hoogheemraadschappen", existing since the 13th century, are organised to manage the water levels. One of these authorities, the “Hoogheemraadschap van Delfland” (Delfland Waterboard or simply Delfland) is in charge of an area comprised by 57 polders covering about 40,000ha.
Delfland is bordered by water bodies to the west (the North Sea) and to the south, namely the Nieuwe Maas and the Nieuwe Waterweg; Figure 3-2), which are at a higher elevation than the land. In order to keep the water level within predefined ranges in the area, each polder drains its extra water into a storage basin through small pumping stations. The storage basin is formed by sets of canals and lakes that are generally higher than the polders, from which six main pumping stations, with a total capacity of 54m$^3$/s, ultimately drain the excess of water either into the North Sea, the Nieuwe Maas or the Nieuwe Waterweg.

The key parameters for the management of Delfland are water level and water quality (Lobbrecht 1997). These variables affect seven well defined interests: flood prevention, ecology, horticulture, pasture agriculture, recreation, navigation and operations. A summary of these interests is shown in Figure 3-3, where their relative priorities are shown by the chart divisions. It is worth mentioning that these divisions were made to be used in the operational optimisation process of the water system and that the values are not determined by the Waterboard.
Figure 3-2. Land uses, main water system components of Delfland region and location of the polders of Pijnacker. Adopted from Lobbrecht (1997).

Figure 3-3. Interest-weighting chart Delfland water system. Adopted from Lobbrecht (1997), p 183.
3.2 Description of the polder system of the Pijnacker region

The region of Pijnacker, with an area of 18.80 km$^2$, is located to the East of the Delfland area. The region is mainly rural, with some urban development (5.7 km$^2$) and glasshouses (2.82 km$^2$). Figure 3-4 shows the land use in the region and Figure 3-5 shows, in a schematic way, a typical profile in the region.

![Figure 3-4. Land use in the region of Pijnacker](image)

![Figure 3-5. Schematic profile of the polders of Pijnacker](image)
The polder system consists of four main polders, namely Polder van Bresland (I), Oude polder van Pijnacker (II), Noordpolder van Delfgauw (III) and Nieuwe of drooggemaaktepolder (IV), each being divided into 6, 63, 27 and 31 smaller polders, respectively (see Figure 3-6). For simplicity, the entire system is referred to as “Pijnacker” throughout this thesis, as this is the name of the most important village in the region.

3.2.1 Composition of drainage units

From the hydrologic point of view, the four major polders and their 127 small polders are hydrologically independent response units, with 28 unique target water levels. The system has 13 pump stations and 21 fixed weirs that are operated in order to keep the water levels in the canal network between limits defined by the water management of the region (Figure 3-7). The arrows indicate the flow directions at the structures, which generally go from West to East and from North to South, with the last destination being the storage basin. The color scale indicates the target of the water levels, which is also an indication of the relief of the region.
3.2.2 Characteristics of the canal network

The main canals of the Pijnacker region have a total length of about 68.4 km and cover a surface of 0.45 km². The geometry of the canal network allows for a maximum storage volume of about 487,000 m³. In Figure 3-8 the storage curve of the canal network is presented for different levels, and shows two well-defined sections in which an increase in storage can be observed. The rate of change of storage drops between the levels -4.5m and -4.0m, because few canals exist with these levels, as can be confirmed by looking at Figure 3-7.
3.2.3 Structures of control

Although there are weirs and pumps in the region, the former are not really operated, but are generally fixed. For this reason, the water levels are controlled by pumps that are operated with simple on/off controllers. The operational on/off levels and the capacities of the 13 pump stations shown in Figure 3-6 are presented in Table 3-1. From Figure 3-6 it can also be observed that pumps 1 and 2 drain the polder I, pumps 4 and 5 drain the polder III, pump 6 drains the polder IV and finally pump 7 drains the polder II into the storage basin. This means that the remaining pumps have the function of raising the water from the lowest parts of each polder.

<table>
<thead>
<tr>
<th>Pump</th>
<th>Capacity (m³/s)</th>
<th>On level (m)</th>
<th>Off level (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.45</td>
<td>-5.1</td>
<td>-5.2</td>
</tr>
<tr>
<td>2</td>
<td>0.64</td>
<td>-2.1</td>
<td>-2.11</td>
</tr>
<tr>
<td>3</td>
<td>0.12</td>
<td>-2.27</td>
<td>-2.32</td>
</tr>
<tr>
<td>4</td>
<td>0.06</td>
<td>-2.9</td>
<td>-2.95</td>
</tr>
<tr>
<td>5</td>
<td>0.53</td>
<td>-2.05</td>
<td>-2.15</td>
</tr>
<tr>
<td>6</td>
<td>1.05</td>
<td>-3.1</td>
<td>-3.11</td>
</tr>
<tr>
<td>7</td>
<td>2.10</td>
<td>-2.65</td>
<td>-2.75</td>
</tr>
<tr>
<td>8</td>
<td>0.49</td>
<td>-3.15</td>
<td>-3.16</td>
</tr>
<tr>
<td>9</td>
<td>0.05</td>
<td>-3.45</td>
<td>-3.5</td>
</tr>
<tr>
<td>10</td>
<td>0.13</td>
<td>-3.15</td>
<td>-3.16</td>
</tr>
<tr>
<td>11</td>
<td>0.60</td>
<td>-5.25</td>
<td>-5.35</td>
</tr>
<tr>
<td>12</td>
<td>0.30</td>
<td>-5.9</td>
<td>-6</td>
</tr>
<tr>
<td>13</td>
<td>1.00</td>
<td>-5.62</td>
<td>-5.72</td>
</tr>
</tbody>
</table>

Although the weirs have the possibility of being operated by changing their crest level, this is not a common practice in the current management of the Pijnacker region.

3.3 Water level monitoring network

At present, automatic water level gauges reporting every 15min are located at the pumping stations for controlling the pumps. Additionally, manual gauge scales are placed in the region, some of them located at the pump stations as a backup reference for operation, and others located elsewhere. The manual gauges are read once a month at places where no automatic gauges exist.

The limitation of the current water level monitoring network is that it is exclusively dedicated to the operation the water system through the control of its pumping stations. However, as every single polder has an associated target water level, the process of knowing the current state of every point in the system is difficult and very expensive. For this reason, the Delfland Water Board may miss out-of-range water levels that affect one or more water users. One way to estimate the state of the water levels at any point in the system, therefore, consists on having a reliable model of the network of canals.
3.4 Description of the model of the Pijnacker polder system

This section describes the model of the Pijnacker polders that was used to develop the approaches described in the chapters 5, 6 and 7. The objectives, topology, discretization and components are summarized in the following paragraphs.

3.4.1 General description of the existing hydrodynamic model

The available model was built and instantiated by the Delfland Water Board between 2005 and 2006, for the purpose of evaluating the state of their system under normal operating scenarios (static and dynamic cases for summer and winter operation in 2006 and system design for 2010) and also under extreme events.

It is a fully 1-D hydrodynamic (HD) model with an attached rainfall-runoff (RR) model that runs in parallel (both models run simultaneously, sharing information at every time-step). Together, both models include 2,692 link elements that connect 3,300 nodes, 83% belonging to components of the RR model, such as friction and boundary elements and areas representing glasshouses, open water, paved, and unpaved areas. The remaining 17% of the nodes represent flow elements of the HD model, such as boundaries, bridges, flow calculation points, flow connections (including those connected to the RR model),

Figure 3-9. Location of the existing water level gauges in the Pijnacker polders.
cross sections, culverts, fixed calculation points, lateral flows, measurement stations, pump stations and weirs. The time step for both HD and RR models is set to $\Delta t=3$ min and the reporting time, also for both models, is 15 minutes. The spatial discretization of the calculation points of the network is 50 m on average.

### 3.4.2 Rainfall-runoff connections

The connections between the HD model and the RR model occur at 130 connections points spread over the canal network. Every single drop-shaped point in Figure 3-10 represents a connection, at which the computation of the hydrological processes occurring at the corresponding areas is carried out. It is assumed that a rainfall event is evenly distributed over the area.

![Figure 3-10. Connection points of the hydrodynamic (HD) model and the rainfall-runoff (RR) model.](image)
In order to analyse the performance, validity and implications of the methods for monitor location developed in Chapters 5, 6 and 7, two water systems of completely different nature are considered in this Thesis. First, the polder system of the region of Pijnacker, The Netherlands, introduced and described in Chapter 3, was selected as being representative of a flat, highly controlled water system. Second, in order to include a natural stream, the Magdalena River, Colombia, is introduced and described in this Chapter.

The following section introduces the general characteristics of the Magdalena River, its catchment, tributaries and wetlands. This is followed by a description of the river users and the different water interests, as well as a discussion of how the performance of the river as a water system can be defined. Next, the 1-D hydrodynamic model development which is the base of the developed methods is presented. Finally, the limitations of the model are listed at the end of the chapter.

### 4.1 Introduction

The Magdalena River, the main river of Colombia, Figure 4-1, runs for about 1,530 kilometres from South to North flowing into the Caribbean Sea, draining a catchment of 273,459 km², equivalent to 24% of Colombia, and where 77% of the population lives (Cormagdalena and ONF_Andina 2007). The mean annual discharge at the river mouth is 7,200 m³/s, with a mean low discharge of 4,068 m³/s in March and a mean high discharge of 10,287 m³/s in November (Restrepo et al. 2006). These figures make the Magdalena the largest river discharging into the Caribbean Sea.

From a morphologic point of view, the Magdalena River is divided into three regions, namely the High Magdalena (682 km, from where the river originates until a zone of rapids nearby the city of Honda), the Middle Magdalena (500km, defined from the Honda’s rapids until the town of El Banco), and the Low Magdalena (430km, defined from Regidor until the discharge into the Caribbean Sea, at the city of Barranquilla) (Julius_Berger_Consortium 1926).

This classification is also convenient for the navigation activity, as each sector needs its own kind of ships to navigate. In terms of the mean hydraulic slope, the High Magdalena
is steep, the Middle Magdalena is moderate and the Low Region is almost flat (see Table 4-1).

Figure 4-1. General location of the Magdalena River and its catchment

Table 4-1. Mean hydraulic slope by sectors, Magdalena River

<table>
<thead>
<tr>
<th>Region</th>
<th>Sector</th>
<th>Mean hydraulic slope (cm/km)</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Magdalena</td>
<td>Embalse de Betania – Purificación</td>
<td>101</td>
</tr>
<tr>
<td></td>
<td>Purificación – Pto. Salgar</td>
<td>45</td>
</tr>
<tr>
<td>Middle Magdalena</td>
<td>Pto. Salgar – Barrancabermeja</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>Barrancabermeja – Regidor</td>
<td>33</td>
</tr>
<tr>
<td>Low Magdalena</td>
<td>Regidor – Banco</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Banco – Magangué</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Magangué – Calamar</td>
<td>6</td>
</tr>
</tbody>
</table>

The area of interest for this research is the middle and low Magdalena. These regions are important for the country not only in terms of economic activities such as navigation, fish production and agriculture, but also because these regions suffer most from flooding (Cormagdalena and ONF_Andina 2007; IDEAM 2001).
4.1.1 Tributaries

The main tributaries, located mainly in the middle reach, have an important influence on the river’s behaviour in terms of discharge. Figure 4-2 presents the main tributaries located in the middle and low sector of the Magdalena River, along with the most important cities.

The Cauca River is the main tributary of the Magdalena. In fact, this river is the second biggest river of Colombia. In the middle reach, the biggest rivers discharging into the
Magdalena are the Carare, Sogamoso, La Miel, Nare and Cimitarra. Although Figure 4-3 shows 21 tributaries, it is worth mentioning that the Magdalena River also receives discharges from an important number of small streams, some of them functioning as connections between the river and the wetlands.

![Figure 4-3. Mean discharges of the main tributaries of the middle and low sector of the Magdalena River](image)

### 4.1.2 Wetlands

The limit between the middle and the low regions is very special because of a number of reasons. First, a relatively sudden slope transition, due to the ending of the mountainous system of the Andes, takes place. Second, the geology of the region, characterized by the dynamics of the Pacific and the Caribbean tectonic plates and the subsequent geologic faults, generates the so-called “Depresión Momposina” (see Figure 4-4), a depression zone in which subsidence process at a rate between 2mm and 4mm per annum (Martínez 1981; Smith 1986; Van der Hammen 1986). Third, the hydraulics of the region, characterized by the discharges of the rivers Cesar, San Jorge and Cauca (the biggest tributary of the Magdalena, which divides into the branches Loba and Mompox; see Figure 4-4), forms a system comprised of hundreds of interconnected “ciénagas” or water bodies that form a natural reservoir which absorbs the peak flows of the river. Finally, the sedimentation processes, characterized by the deposition of 21 Mt/yr, yielding a sedimentation rate between 2 mm/yr (Restrepo 2008) and 3mm/yr (Van der Hammen 1986), which implies that the net elevation change of the wetland bed is barely visible. Additionally, it is important to mention that the Momposina depression is an area with a very important biodiversity in fauna and flora and with a high potential for fish, forest and crop production (Aguilera 2004; Aguilera 2009). However, this biodiversity is under risk because of uncontrolled human interventions and the lack of protection policies (Galvis and Mojica 2007; Múnera et al. 2004; Muñoz et al. 2003; Naranjo et al. 1999). A comprehensive description of the region can be found in DNP, FAO et al. (2003).
Although it is known that the Momposina depression is one of the biggest inner deltas in the world (Van der Hammen 1986), there is no clarity on its extension. Restrepo (2005) states that this “tectonic tray” has an approximate area of 1850 km$^2$, while in another publication, the same author indicates an area of 800 km$^2$ (Restrepo et al. 2006). In terms of the area occupied by the wetlands, the Julius Berger Consortium (1926) estimated, with instruments of the time, an area of “at least” 2,200 km$^2$. In this study, the area of the wetlands in the Momposina depression, taken from El Banco until 25 km to the North of Plato, is 3,096 km$^2$, which includes the Mojana region and the Zapatosa wetland (365 km$^2$ alone). These estimates come from an analysis of the SRTM Water Body Data (SWBD) data set.

![Figure 4-4. Inner delta and wetlands of the Momposina depression](image)

### 4.1.3 Description of the existing Monitoring Network

Nowadays, more than 3,300 hydro-meteorological stations of different types are operating in the Magdalena – Cauca catchment. The methods presented in this thesis for the case study, however, consider only the water level monitoring network.

The existing water-level gauges for the river were placed initially to support decision-making concerning local problems in the main populated areas, related to flood control and navigation, while keeping operation and maintenance costs low. However, from a
global perspective, the information collected by these gauges is limited to supporting decision and policy-making for navigation, flood control and other issues at other points of the river.

Figure 4-5 shows the location of the existing limnigraphs and limnimeters to monitor the water levels of the middle and low Magdalena River, and the present age of their records. It can be observed that Salgar, Berrío, Barrancabermeja, Wilches, El Banco and Calamar are the oldest stations and therefore these are always included in all the navigation and flooding studies in the Magdalena River.

Figure 4-5. Age of the existing limnigraphic and limnimetric stations in the middle and low Magdalena River (based on the year 2010)
During the data collection for this thesis, a number of informal interviews were carried out with different people involved for years in several studies about the Magdalena River. They described the following common problems (some also found in the literature) with the water level monitoring network of the Magdalena River:

- The distance between two consecutive gauges with sufficient historical records is huge (157km between Salgar and Berrío, 105km between Berrío and Barrancabermeja, 676km between Barrancabermeja and El Banco and 309km between El Banco and Calamar). Even the most recent, intermediate gauges are still too far apart to draw proper conclusions about water levels in the Magdalena River (Cormagalena 2004).

- Some gauges have changed been resited in time, because of a number of factors that range from accidents with ships to incidents with high flows that wipe out the gauges, and situations that have not been properly reported (Cormagalena 2004).

- The gauges are read by human observers who must record the levels and report them in a predefined format. Although there is no documented evidence, situations like forms being filled without actually looking at the gauges were mentioned, which add enormous uncertainty to the quality of the data. Other problems like illiterate or non-committed observers and vandalism were faced by even earlier studies in the Magdalena River (see e.g., Julius_Berger_Consortium 1926).

- The majority of the gauges have a relative reference level, which means that they do not provide the absolute water level in a unique elevation system. This situation has at least three causes: first, the institution in charge of placing, maintaining and operating the gauges (Institute for Hydrology, Meteorology and Environmental Studies, IDEAM) is different from the institution in charge of providing maps, levels and the topographical information of the country (Geographic Institute Agustín Codazzi, IGAC), and is also different from the institution in charge of making the decisions about the floods and navigation on the river from a global perspective (Corporación Autónoma Regional del Río Grande de la Magdalena, CORMAGDALENA). Second, the gauges were initially installed with relative or arbitrary references, enough to make decisions of a local nature. Third, there exists little or no interest in georeferencing the existing gauges to a unique, precise reference system, as can be inferred from the persistently ignored recommendations of old studies (Julius_Berger_Consortium 1926; Mitch 1973) as well as relatively new studies (see e.g., Cormagalena 2000b; Cormagalena 2004). Yet an additional reason is that IDEAM, which is the entity in charge of the hydrological forecasting in the country, has been using statistical methods that do not require absolute but relative levels (IDEAM 2005). Indeed, a number of studies to forecast water levels for navigation purposes in the Magdalena River have used relative water levels (see e.g., Dominguez et al. 2009; Fernandez et al. 2010; IDEAM 2005; Mitch 1973; Rivera et al. 2004; Rojas 2006).
• The gauges with absolute datum have actually been referenced using the bench marks of the closest road projects, which, in turn, are barely referenced to the same absolute datum. This means that the absolute datum of these gauges can only be considered to be approximate.

• Some authors consider that the huge amount of human intervention in the river during the last decades, such as dredging for navigation purposes, the construction of dikes for flood control, and the illegal (and therefore not documented) closure of the natural wetland-river connections for land reclamation, have necessarily affected the time series records of the river, and for this reason any statistical analysis of them should be done with care (see e.g., Cormagdalena and Fedenavi 2007).

• Additionally, the daily discharge data at some stations is based on rating curves, so that every (relative) water level record registered at the gauge is converted to a discharge. Although this is a common, accepted practice, several issues added uncertainty to these curves. For instance, the changing morphology of the Magdalena River implies that the sections where these rating curves have been deduced have experienced important modifications. For this reason, some authors have decided to eliminate old records, in order to guarantee in some way a constant hydraulic section (see e.g., Cormagdalena 2006; Cormagdalena and Fedenavi 2007).

4.2 Performance of the Magdalena River

The main water interests of the river, which are in permanent conflict, are navigation, agriculture and pasture, flood control and fish production. Navigation in the middle and low Magdalena River is important because it is a way to connect the significant/strategic, productive cities of the interior with the ports of Barranquilla and Cartagena facilitating the trading activities on the Caribbean Sea and the Atlantic Ocean. Naturally, low water levels due to either dry periods or sediment deposits imply economic losses for ships that get stuck in bar sands along the river. A comprehensive review of the navigation in the low Magdalena River can be found in Alvarado et al. (2008).

The social and political situation of the people settled in the Magdalena River may explain the type of decisions concerning the management of the river. A well known situation is that the land is generally owned by a few powerful families that exploit it for agriculture and cattle (Aguilera 2004; DNP et al. 2003). In order to get more land, these people started land-reclamation activities, such as closing the streams that connect the river and the wetlands. Other activities included the construction of carriage ways for transportation purposes, which help to close the stream connections. These practices have affected the ecology of the region negatively, together with fish production. Detailed environmental impacts in the middle Magdalena River due to human activities can be found in Cormagdalena (2007) and DNP et al. (2003).

From a flood control perspective, the solutions have considered mainly structural methods such as lateral dikes along the river. Although this is not a good practice,
(blocking the floodplain decreases the channel conveyance and increases flood stages; also their construction eliminates natural storage originally available on the flood plain, generating an increase of the runoff concentration and of the flood peaks downstream), between 2004 and 2009 a total of 517 km of linear concrete walls, piles and dikes have been built along the river (Cormagdalena 2009), which might make the situation worse in the future. Non-structural methods such as warnings are prepared at national level by the Institute for Hydrological, Meteorology and Environmental Studies (IDEAM), for local emergency associations to prepare affected communities for possible flood occurrences. Although efforts to improve the quality of these warnings are carried out, at present they still are very fuzzy.

Finally, fishing production has dramatically dropped. Some authors explain this reduction as a result of agricultural, urban, and industrial development and deforestation in the river's watershed (see e.g., Abramovitz and Peterson 1996; Silva; Valderrama and Zarate 1989); others think that the reason is due to over-exploitation of fisheries by displaced people (see e.g., Galvis and Mojica 2007); however, local reports made in the field with the fishermen (see e.g., Campo 2001; Rodriguez 2001; Romero 2001), conclude that the drop in fishing production is due to the reduction of areas and depths of the wetlands by human intervention in the river-wetlands connections. A comprehensive diagnosis of the fishing problems in the Magdalena River can be found in Gualdrón (2006).

As mentioned above, a good performance of the Magdalena River can be defined by the extent to which the navigation, agriculture and cattle, flooding and fish production fulfil the expectations of each water user.

### 4.3 Development of the hydrodynamic model for the Magdalena River

The theoretical approaches for the design and evaluation of monitoring networks presented in Chapters 5 and 6 require the use of a hydrodynamic model. For this reason, the hydrodynamics of the middle and low part of the Magdalena River have been modelled. In this section the details of this model are presented. It is worth mentioning that the developed model does not include a hydrological component, because the analysis of the river response under rainfall events is beyond the objectives of this thesis. However, the developed model may be updated and complemented for other uses.

#### 4.3.1 Data used

The data available to build and instantiate the hydrodynamic model of the Magdalena River, between Salgar and Calamar, include information of stages and discharges of river stations, river network and bathymetry obtained for different studies and satellite floodplain elevations. The details of the data used are described as follows.

**Stage and discharge**

Daily, multiannual water level and discharge records are available for the low and middle Magdalena River and its tributaries range from 1974 to 2003. However, the most complete data for the tributaries and including 15 out of the 32 existing stations (see
Figure 4-5) are for the year 1995 (see Table 4-2 and Figure 4-6) and therefore the model has been built for this period.

![Diagram of hydrologic data records](image)

Figure 4-6: Available hydrologic data records of discharges (Q) and water levels (h) at river stations for 1995.

<table>
<thead>
<tr>
<th>Station</th>
<th>Discharge</th>
<th>Source*</th>
<th>Stage</th>
<th>Datum</th>
<th>Source**</th>
</tr>
</thead>
<tbody>
<tr>
<td>Magdalena River</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Salgar</td>
<td>365</td>
<td>R</td>
<td>365</td>
<td>165.9</td>
<td>A</td>
</tr>
<tr>
<td>Berrio</td>
<td>365</td>
<td>R</td>
<td>365</td>
<td>104.5</td>
<td>A</td>
</tr>
<tr>
<td>Barranca</td>
<td>365</td>
<td></td>
<td>365</td>
<td>70.5</td>
<td>A</td>
</tr>
<tr>
<td>San Pablo</td>
<td>261</td>
<td>R</td>
<td>261</td>
<td>54.6</td>
<td>A</td>
</tr>
<tr>
<td>Regidor</td>
<td>365</td>
<td>D</td>
<td></td>
<td>23.5</td>
<td>B</td>
</tr>
<tr>
<td>Peñoncito</td>
<td>365</td>
<td>D</td>
<td>263</td>
<td>16.6</td>
<td>B</td>
</tr>
<tr>
<td>El Banco</td>
<td>321</td>
<td>R</td>
<td>365</td>
<td>19.6</td>
<td>A</td>
</tr>
</tbody>
</table>

Table 4-2. Number of days of 1995 with discharge and stage data and datum of gauges
**Chapter 4 - Case study 2: Magdalena River**

<table>
<thead>
<tr>
<th>Location</th>
<th>Rating Curve</th>
<th>Discharge (m³/s)</th>
<th>Data Reported</th>
<th>Datum (masl)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tacamocho</td>
<td>D</td>
<td>365</td>
<td>297</td>
<td>5.62</td>
</tr>
<tr>
<td>Plato</td>
<td></td>
<td></td>
<td>271</td>
<td>2.85</td>
</tr>
<tr>
<td>Tenerife</td>
<td></td>
<td></td>
<td>271</td>
<td>0.19</td>
</tr>
<tr>
<td>Calamar</td>
<td>R</td>
<td>365</td>
<td>365</td>
<td>-0.20</td>
</tr>
</tbody>
</table>

**Mompox branch**

<table>
<thead>
<tr>
<th>Location</th>
<th>Rating Curve</th>
<th>Discharge (m³/s)</th>
<th>Data Reported</th>
<th>Datum (masl)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Santa Ana</td>
<td>D</td>
<td>365</td>
<td>365</td>
<td>10.92</td>
</tr>
<tr>
<td>San Roque</td>
<td></td>
<td></td>
<td>365</td>
<td>15.73</td>
</tr>
</tbody>
</table>

**Loba branch**

<table>
<thead>
<tr>
<th>Location</th>
<th>Rating Curve</th>
<th>Discharge (m³/s)</th>
<th>Data Reported</th>
<th>Datum (masl)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Armenia</td>
<td>D</td>
<td>365</td>
<td>365</td>
<td>12.15</td>
</tr>
<tr>
<td>Magangue</td>
<td></td>
<td></td>
<td>365</td>
<td>9.241</td>
</tr>
</tbody>
</table>

*R: Rating curve, D: data reported by IDEAM (Cormagalena 2006); **Datum (masl) reported by A: Cormagalena’s website; B: studies by LEH-UN; C: studies by LEH-LF, Uninorte.

**Boundary conditions**

The boundaries of the model are the 1995 discharges at Puerto Salgar (upstream) and the 1995 water levels at Calamar (downstream). Additionally, the time series of the 12 main tributaries of the river in the considered sector (see Figure 4-6 and Figure 4-7) were included as point sources.

![Figure 4-7. Mean discharge of tributaries of the Magdalena River and mean discharges for the year 1995, used as model inputs.](image)

**River network and bathymetry**

The model of the Magdalena, composed by the branches Magdalena (Loba) and Mompox, is built using a network point every 200 meters. In total, the Magdalena branch contains 4058 points, with chainages ranging from K906+000 at Puerto Salgar (upstream) to K94+600 at Calamar (downstream). Similarly, the Mompox branch, formed by 1080 points, connects to the Magdalena network at the chainage K402+600 upstream (just after El Banco) and at K227+800 downstream. The maximum distance between two adjacent computational points is $\Delta x = 5000m$. 
The bathymetric information used for the hydrodynamic model were obtained during works made during 1999 and 2000 for the navigation studies ordered by Cormagdalena, the government’s institution dedicated to the management of the river. Two main institutions, namely the LEH-LF (del Norte University, Barranquilla) and the LEH-UN (National University, Bogotá) were in charge of the execution of the studies, the former in charge of the low Magdalena (Cormagdalena 2000b) and the latter of the middle Magdalena (Cormagdalena 2000c); together they released the first Magdalena’s River Navigation Booklet (Cormagdalena 2000a). Yet another navigation booklet was produced in 2004, and a project to obtain frequent bathymetries for navigation assistance in near real-time (Alvarado 2006). However, the works of the year 2000 were selected in order to use satellite elevation data from HydroSheds (HS) by Lehner et al. (2006) to complement the sections. HS is a hydrologically-corrected version of the of the Shuttle Radar Topography Mission (SRTM) elevation data obtained by the Space Shuttle Endeavor mission (National Aeronautics and Space Administration, NASA) in February of 2000.

Cross sections
As the primary focus of field work is navigation, the information collected was insufficient for the modelling process. Certainly, the work does not include the flood plain topography so cross sections are incomplete. Additionally, the bathymetries are expressed in terms of depths in order for the pilots to know how much load they may transport without getting grounded in the river.

Unfortunately, only local projects such as river bank protections or flood control works have information about flood plain topography. This is because the local nature of these projects makes it unnecessary to reference the elevations to an absolute datum.

As stated above all the cross sections available for modelling are limited to the main channel of the river and therefore the elevation of the embankments is not well defined. For this reason, it was decided to combine the bathymetry information of the year 2000 to the HS DTM data, in order to obtain complete cross sections for the model. However, in cases where there was not enough information or where the available bathymetry did not have a reliable elevation transformation, the bathymetry obtained during other years was used. The general procedure to combine the two sources of information is described as follows:

- Select the place where a cross section is needed.
- Search for the corresponding water level at the point where the cross section is located and at the closest referenced gauge during the period of the field work. If the data of the referenced gauge is not available, replace it by the water level that is exceeded 50% of the time for the date and year under consideration.
- Estimate the absolute elevation reference of the cross section, based on the reading of the closest referenced water level gauge and the hydraulic slope during the field work.
- Draw in plan view the cross section line of the bathymetry and extend it for at least 100 to 500 meters depending on the characteristics of the river.
Chapter 4 - Case study 2: Magdalena River

- Extract the profile from the HS DTM using the line obtained in 4.
- Replace the corresponding points of the HS DTM by the bathymetry points.

This procedure was used to produce 33 cross sections that were included in the model. An example of the obtained cross section is shown in the Figure 4-8, for the case of a cross section near La Dorada – Puerto Salgar.

![Figure 4-8. Example of a composite cross section near La Dorada - Puerto Salgar](image)

**Wetlands**

At present, the hydrologic and hydraulic information about the wetlands is limited to what can be inferred from satellite images, aerial photography and cartography, which is basically the extent of the water bodies areas. Perhaps the unique attempt to establish a water balance in part of the considered region is the work by Díaz-Granados et al. (2001), in which the wetlands of the Mojana Region and the Zapatoza, downstream the Cesar River, were modelled. However, the authors recognize that their efforts provide only a “qualitatively valid approximation” of the behaviour of the wetlands, due to the deficient amount of information, especially regarding the elevation data, a key component that drives the flow pattern in such a flat area. Additionally, during the information collection for this thesis, no bathymetry information for the wetlands was found.

From a modelling point of view, the wetlands at the limit between the middle and low sector of the Magdalena River were simulated using control structures. For this purpose, the existing wetlands were grouped into four large reservoirs that were taken into consideration in the model (see Figure 4-9). These reservoirs were located at chainages K516+800, K403+800, K381+800 and K269+800 from upstream to downstream, respectively. The main idea behind the control structures is that during the wet seasons, the excessive flow coming from Magdalena partially discharges into the wetlands through weirs. Similarly, during the dry season, allow water flow from the wetlands to the river. Details of the four grouped wetlands are presented in Table 4-3.
Optimisation of monitoring networks for water systems

The elevations and the areas presented in the Table 4-3 were obtained from GIS analysis of the HS DTM data, which is still an approximation. It is clear that due to the deficient cartography and elevation information, the exact connections between wetlands and between rivers and wetlands is unknown, and therefore the real behaviour of the wetlands in terms of storage volumes and peak flood attenuation of the rivers Magdalena and Cauca is uncertain, so further investment in monitoring and research in general is needed (DNP et al. 2003).

4.3.2 Simulation characteristics

The data available allows the user a simulation period from 01/01/1995 at 12:00:00 PM to 12/31/1995 at 12:00:00 PM. With a calculation time step of 10 minutes and provided $\Delta x = 5000m$ and $h = 3m$, the Courant number for stability and accuracy criteria yields

$$Cr = \frac{\Delta t}{\Delta x} = \frac{\sqrt{gh} \times 3 \times 10 \times 60}{5000} = 0.65 < 1.0$$
Table 4-3. Characteristics of the grouped wetlands

<table>
<thead>
<tr>
<th>Wetland</th>
<th>Description</th>
<th>Elevation (masl)</th>
<th>Area (ha)</th>
</tr>
</thead>
<tbody>
<tr>
<td>W1</td>
<td>Accounts for wetland at the discharge of the Lebrija River, Ciénaga Simití and the diversion of the Morales branch</td>
<td>38</td>
<td>37,414</td>
</tr>
<tr>
<td></td>
<td></td>
<td>42</td>
<td>37,414</td>
</tr>
<tr>
<td>W2</td>
<td>Wetland2. Accounts for the Zapatoza wetland (nearby El Banco) and other water bodies at the left shore of the Magdalena River, in front of El Banco</td>
<td>23</td>
<td>2,473</td>
</tr>
<tr>
<td></td>
<td></td>
<td>24</td>
<td>59,864</td>
</tr>
<tr>
<td></td>
<td></td>
<td>25</td>
<td>78,551</td>
</tr>
<tr>
<td></td>
<td></td>
<td>26</td>
<td>102,778</td>
</tr>
<tr>
<td></td>
<td></td>
<td>27</td>
<td>109,411</td>
</tr>
<tr>
<td></td>
<td></td>
<td>28</td>
<td>113,081</td>
</tr>
<tr>
<td></td>
<td></td>
<td>29</td>
<td>116,852</td>
</tr>
<tr>
<td>W3</td>
<td>Wetland3. Accounts for the wetlands of the Mojana region, at both sides of the discharge of the Cauca River into the Magdalena River</td>
<td>10</td>
<td>31,858</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12</td>
<td>55,011</td>
</tr>
<tr>
<td></td>
<td></td>
<td>13</td>
<td>60,571</td>
</tr>
<tr>
<td></td>
<td></td>
<td>17</td>
<td>85,054</td>
</tr>
<tr>
<td></td>
<td></td>
<td>19</td>
<td>126,217</td>
</tr>
<tr>
<td>W4</td>
<td>Wetland4. Accounts for the wetlands of the Momposina Depression</td>
<td>10</td>
<td>43,721</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11</td>
<td>74,600</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12</td>
<td>99,335</td>
</tr>
<tr>
<td></td>
<td></td>
<td>14</td>
<td>108,833</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15</td>
<td>122,234</td>
</tr>
</tbody>
</table>

4.3.3 Model calibration

Once the model was instantiated, the first runs were aimed to check the continuity of the discharges in the river to get the correct volume of water, compared to the flow measurements given at different points. As a first attempt, no wetlands were included. Due to the missing data of some tributaries before 30 April 1995, the analysis was carried out from 1 May 1995. The first results show that at Puerto Berrío the modelled discharges
replicate the measurements very well, but the results become less acceptable in the downstream stations. At Regidor, for instance, although the general trend of the discharge curve is acceptable (see Figure 4-10(a)), the shape of the measurement curve is smoother than the modelled one; also the modelled volume is persistently higher (between 700 and 1000m$^3$/s that is not reflected in the measurements). At Calamar, additionally, the flow curve is completely different in terms of trend, quantity and shape; see Figure 4-10 (b).

Figure 4-10. Modelled and measured discharges at Regidor (a) and Calamar (b), first check

Figure 4-11. Modelled and measured discharges at Regidor (a) and Calamar (b), second check
Although missing inflows may indicate that other minor reaches are not included or that rainfall over the river may have a large influence, the deficient quality of the rating curves used to produce the discharge time series could also be an important source of error. In order to account for the missing inflows, lateral inflows were added into three sections of the Magdalena River: from Berrío to San Pablo (400m³/s), from San Pablo to Regidor (800m³/s) and from Magangue to Tacamocho (500m³/s). The results of these inflows are shown in Figure 4-12.

For the Mompox branch, the results are presented in Figure 4-13.
It is evident that the modelled discharge at Regidor replicates best the measurements with the new lateral inflows. However, at the downstream boundary (Calamar) neither the shape of the curve nor the discharges are reproduced well. At this point, therefore, the necessity of involving the wetlands is evident. On the one hand, the regulatory nature of the wetlands compensates for the water balance downstream, and on the other hand their presence attenuates the discharge in time, making the time series smoother just as they are observed at the downstream stations.

After the acceptable reproduction of the volume of water at different points of the river, the measured and modelled water levels were adjusted by changing the roughness coefficients. The final Manning coefficients are presented in the Table 4-4. The coefficients for the intermediate places were estimated using linear interpolation between two consecutive stations.

<table>
<thead>
<tr>
<th>River Name</th>
<th>Station Name</th>
<th>Chainage (m)</th>
<th>Resistance Number (local values)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Magdalena</td>
<td>Puerto Berrío</td>
<td>759600</td>
<td>0.022</td>
</tr>
<tr>
<td>Magdalena</td>
<td>Barrancabermeja</td>
<td>656600</td>
<td>0.048</td>
</tr>
<tr>
<td>Magdalena</td>
<td>San Pablo</td>
<td>607800</td>
<td>0.038</td>
</tr>
<tr>
<td>Magdalena</td>
<td>Magangué</td>
<td>258600</td>
<td>0.025</td>
</tr>
<tr>
<td>Magdalena</td>
<td>Tacamócho</td>
<td>212200</td>
<td>0.025</td>
</tr>
<tr>
<td>Magdalena</td>
<td>Plato</td>
<td>167200</td>
<td>0.040</td>
</tr>
<tr>
<td>Mompox</td>
<td>San Roque</td>
<td>39300</td>
<td>0.025</td>
</tr>
</tbody>
</table>

Although the water level is not completely well reproduced by the model, the trend and values follow the pattern of the measurements, as can be observed in Figure 4-14 for the Regidor station.

A more detailed description of the developed model and the calibration process can be found in He (2009), a Master of Science thesis that supports the present dissertation.
4.4 Limitations of the model

Although a big effort was made to find a model that reproduces the measurements within acceptable ranges, there is a high uncertainty in a number of inputs that have been discussed in this chapter and these are summarized below.

Unreferenced water levels
The well-known problem of the absolute reference datum for elevation is perhaps the main source of uncertainty in the model, because it affects not only the direct topology of the model such as the bathymetry and the cross sections, but also the boundary conditions. Additionally, this also affects the measurement time series, used to evaluate the performance of the model, which require the use of rating curves to transform registered discharges to referenced water levels.

Cross sections
The cross sections used in the developed model have two sources of uncertainty: first, the elevation becomes worse at points that are in the middle of two consecutive stations some distance apart; second, the limited cross section data includes only the main channel of the river. The method designed to complement the cross sections using HS data to describe the floodplains needs to be checked in the field before being used further. Of course, the ideal solution would be to carry out topography surveys that include complete, referenced cross sections.

Rating curves
The discharge data that comes from the rating curves at different stations is an uncertain input, on the one hand, because of the vagueness of the zero-level of the gauges (affected also by the frequent change of their position due to external factors) and, on the other hand, because of the major modifications (natural and human) to the river’s morphology, with the consequent effects on the cross sections where the flow measurements are made (Cormagdalena 2006).

Wetlands
The modelling of the effect of the wetlands on the Magdalena River hydrodynamics is limited because of the number of assumptions made, which include the tentative location of the connections between the wetlands and the river and their capacity, the volume that the wetlands are able to store and, the lack of knowledge of the elevation of each water body, which restricts the analysis of water balances and the dynamics of the region in general. The crest levels of the weir structures used to simulate the wetland-river flow interchange were adjusted until satisfactory shapes of the discharge curves at the stations located in the branches Loba and Mompox and also at the downstream stations were obtained.

As a proper validation of the model was not carried out, it is recommended to perform it by using recent data sets.
Chapter 5
Information Theory for monitor location

A review on Information Theory is presented in Chapter 2, in which the philosophy behind the concept of information, the approaches to measure it and some of its applications are described. This introduction, together with the description of the polder system in the Pijnacker region in The Netherlands in Chapter 3 and Magdalena River, Colombia, in Chapter 4, are the foundation of the methods for determining monitor locations developed in this Chapter.

This Chapter begins with an introduction that includes the main considerations that are common in the development of the methods, namely the use of models as data generators and the description of the estimation of probabilities for the information-related measures. Next, the developed methods are presented in three sections, each one presenting a description of the developed methods, a brief recall of the case study under consideration, the presentation and discussion of the results, and the corresponding conclusions.

The first section introduces an approach for locating water level gauges for monitoring in polders using Information Theory with pairwise criteria for dependency estimation. The case study is in the Pijnacker region. The same problem is addressed in the second section, in which additional practical considerations are included. The monitor location problem is posed as a multi-objective optimization problem and solved with an evolutionary optimization method. The third section considers the problem of locating discharge monitors in the Magdalena River, and applying a multi-objective optimization method and also a rank-based greedy algorithm. The Chapter finishes with a summary of the general conclusions obtained.

5.1 Introduction

Two main considerations characterize the methods presented in this Chapter, namely, the use of models to generate series at each computational point in the water system, and the discretization of the generated time series to perform the probability calculations. Both of these considerations are described in the following sections.
5.1.1 Use of models as artificial data generators

The use of models is significant because they are adopted to replicate the real world in such a way that the states of the system can be reproduced. The models are then used to generate time series from which the information-related quantities are estimated. One of the reasons we use models instead of empirical measurements is so that, in order to do the analysis, we have access to a dense set of points which are needed to obtain a complete picture of the behaviour of the system in terms of the information associated with it. The available measurements are generally limited to a few points making them insufficient to analyse and draw conclusions.

It is considered, therefore, that every calculation point within a model is, in principle, a potential location for a monitoring point within the water system. For this purpose, the methodologies consider the use of hydrodynamic and rainfall-runoff models to generate a water level time series at a dense, finite set of calculation points. In this way, a number of water level records are generated with a predefined record length, from which the information-related measures are calculated.

It is important to note that the solutions are related to the time step at which the data records are generated by the model. Therefore, the model must be manipulated according to the final aim and use of the monitoring network, by considering the proper time and space scales. Similarly, the resultant monitoring network will be adequate for capturing the information content of the physics of the runoff process associated with the rainfall event used to produce the water level time series. In other words, every single rainfall event has associated with it an optimal monitoring network, because different control strategies to operate the system take place (e.g., different pump stations will start pumping at different times and for different periods).

For the case of the Pijnacker (Chapter 3) the water system has been modelled by the Water Board Delfland to make operational decisions under several scenarios. For the case of the Magdalena River, the hydrodynamic model has been instantiated and calibrated within the framework of this thesis (see Chapter 4).

5.1.2 Estimation of probabilities for the calculation of IT quantities

The probabilities required for the estimation of entropy and mutual information are calculated using the well-known histogram based technique, as described for example by Steuer et al. (2002); in this way the choice of a probability distribution to fit the continuous data is avoided. Although there exist a number of nonparametric methods to estimate mutual information (see e.g., Moon et al. 1995; Sharma 2000), this method uses bins as an opportunity to take into account water management issues. The subjective determination of the bin size for the histogram construction (Ruddell and Kumar 2009) is addressed here using the quantization method introduced below.

5.1.2.1 Quantization

The quantization concept comes from the theory of communication systems, and aims to convert an analogous (i.e., continuous) sign into a discrete pulse, in order to allow its
digital transmission by applying the mathematical floor function (denoted as $\lfloor \bullet \rfloor$). The conversion of an analogous value $x$ to a quantized value $x_q$, which is rounded to the nearest multiple of $a$, is performed by:

$$x_q = a \left\lfloor \frac{2x + a}{2a} \right\rfloor$$

(5-1)

The relationship between the bin-size $b$ and the parameter $a$ is given by the quotient of the difference between the maximum and the minimum of the time series and the value $a$.

For the context of this Chapter, water level time series are transformed through Equation (5-1) into “pulses” of discrete information, which produces a regular discretization of water levels (in terms of levels) at irregular intervals of time. An advantage of the approach is that the quantized water level series are “noise-free”, in the sense that high-frequency, low-amplitude water level changes (i.e., dynamic waves) generated by neighbouring pumping stations (see for example the time series in Figure 5-2), are filtered out. Consequently, the value of $a$ can be seen as the minimum dimensional unit of water level for which the management of the system becomes critical. This is crucial when computing Equation (2-1), since high-frequency water level signals give very high values of entropy, but do not necessarily provide information content for water management decision-making. It must be noted that even though the quantization alters the water volume at a point, this is not important since only probabilities of the occurrence of the values are taken into consideration for the entropy calculations.

In order to show how the results may change because of the selection of the parameter $a$, a sensitivity analysis is included at the end of each approach presented in the following sections.

5.2 Information theory-based approach for location of monitoring water level gauges in polders

An approach called Water Level Monitoring Design in Polders (WMP) for locating and evaluating water level gauges in a water-system composed of a highly controlled canal network is presented here. It consists of five parts: 1) the generation of a time series at a very dense set of calculation points using a hydrodynamic model; 2) a quantization method to “clean” the noise from the time series; 3) the use of three different pairwise criteria to evaluate dependency using squared matrices in a similar fashion to Mogheir and Singh (2002); 4) a procedure to locate the gauges following a method similar to Krstanovic and Singh (1992a,b), which aims to find the set of points which together provide the highest information content and are at the same time the most independent of each other; and 5) the evaluation of the multivariate dependency with Equation (2-8), using the grouping property of the Total Correlation (Fass 2006; Kraskov et al. 2003) in order to establish comparisons between the gauges. The first criterion of the procedure in 4) is evaluated with Equation (2-1) and the second is evaluated by three different pairwise methods: transinformation (Equation (2-3)), $DIT_{XY}$ and $DIT_{XX}$ proposed by Yang and Burn (1994) (Equation (5-2) and Equation (5-3)).
Optimisation of monitoring networks for water systems

\[ DIT_{x,y} = \frac{I(X,Y)}{H(X)} \]  

\[ DIT_{y,x} = \frac{I(X,Y)}{H(Y)} \]  

Although the concept of Transfer Entropy (Schreiber 2000) and its application in monitoring design (Ruddell and Kumar 2009) promotes a new pairwise dependency criteria, the dynamic analysis of the time series is beyond the scope of this method.

5.2.1 Description of the WMP methodology

The Water Level Monitoring Design in Polders (WMP) methodology considers two different conditions when locating the gauges: a) the monitors must be as independent as possible from each other (that is, have a low pairwise value); b) the monitors must provide, individually, the highest information content (that is have a high entropy). The procedure is explained as follows:

a) Read and quantize the water level time series generated by the hydrodynamic model for each of the calculation points \( s_i \) (\( i = 1, 2, ..., n \), where \( n \) is the number of calculation points). Each point has an associated sequence of values \( X_i \).

b) Calculate the marginal entropy \( H(X_i) \) for each \( s_i \) with Equation (2-1), from which the values to fulfil condition a) will be taken.

c) For each \( s_i \), calculate the mutual information in Equation (2-3) with respect to each of the remaining points and build the symmetric matrix \( T \), in which \( I(X_i; X_j) \) is equivalent to \( H(X_i) \) (Cover and Thomas 1991)

\[ T = \begin{bmatrix}
    I(X_1; X_1) & I(X_1; X_2) & \cdots & I(X_1; X_n) \\
    I(X_2; X_1) & I(X_2; X_2) & \cdots & I(X_2; X_n) \\
    \vdots & \vdots & \ddots & \vdots \\
    I(X_n; X_1) & I(X_n; X_2) & \cdots & I(X_n; X_n) 
\end{bmatrix} \]  

(5-4)

In this way, the point \( s_i \) will have an associated vector of mutual information \( v_i \) defined by the \( i \)th row (or column) of \( T \). The values to fulfil condition b) will be taken from this matrix.

d) The first monitor \( m_1 \) is located at the point that provides the highest information content of the system, (i.e., the point with the highest entropy value), so \( m_1 = \max (H(X_i)) \).

e) Add \( m_1 \) to the matrix of the monitoring points \( M \).

f) Recover the mutual information vector \( v_i \) of the monitor \( m_1 \):

\[ v_i = I(X_i; X_{m_1}), i = \{1, 2, ..., n\} \]
g) The system can then be divided into two sets of points with respect to their dependency on $m_i$: those that are dependent and those that are independent ($S_{m_i}^{\text{ind}}$).

The second monitor $m_2$ must be selected from the latter in order to fulfil condition a). The set $S_{m_2}^{\text{ind}}$ is obtained by looking at the elements of $v_1$ such that $I(X_i;X_{m_2}) < e$, $e$ being the value of transinformation between $X_i$ and $X_{m_2}$ that is insufficient for the pair to be considered dependent.

h) To fulfil condition b), $m_2$ must have the highest entropy possible of the set $S_{m_2}^{\text{ind}}$ so, $m_2 = \max\left(H\left(X_i \in S_{m_2}^{\text{ind}}\right)\right)$.

i) Recover the mutual information vector $v_2$ of the monitor $m_2$:

$$v_2 = I\left(X_i;X_{m_2}\right), i = \{1,2,...,n\}$$

j) The next monitor $m_3$ must be selected in a similar way, but now using a modified set of independent points $S_{m_3}^{\text{ind}}$ given by the common set of independent points in the overlapping transinformation vectors for the previously selected monitors $m_1$ and $m_2$. Therefore, $v_3 = v_i + v_2$.

k) Set $v_i = v_3$ and the procedure is repeated from step f) until a maximum number of points is reached or until $m_i$ does not provide a significant information content for the remaining system (i.e., its marginal entropy is too low). The matrix $T$ in step c) is replaced by $DIT_{xy}$ and $DIT_{yx}$ when evaluating the DIT-based criteria.

Although the general scheme of the method is based on the studies mentioned in the previous section, several changes are proposed to make it applicable to highly controlled water systems. First, the time series of water level are generated by a hydrodynamic model at a very dense set of computational points, each of which is a potential gauging site. This avoids the use of empirical, historical measurements, which are not available at the required density (and at all significant points in a highly controlled polder system). Second, noisy time series produced by the operation of pumps are filtered, by introducing the quantization concept. Third, the independent sets of time series are defined in a new way using the common set of independent points in the overlapping transinformation vectors for the previously selected monitors. The solutions obtained by each of the pairwise criteria are then evaluated with the concept of Total Correlation to check the independency among monitors and the joint information provided by the set.

### 5.2.2 Case study: Pijnacker region, The Netherlands

The case study is located in a low-lying region of Pijnacker, Delfland, The Netherlands, which has an area of 18.80 km$^2$, 15 pump stations and 21 fixed weirs that are operated in order to keep the water levels between limits defined by the water management of the region. A detailed description of the area is presented in Chapter 3.
In order to apply the WMP approach described above, the following actions are taken. The water level time-series are generated by a hydrodynamic model built by the Delfland Water Board to make operational decisions regarding the control structures under several scenarios. A dense set of $n=1520$ calculation points separated along the canals by a distance of 15m on average is used with the rainfall event shown in Figure 5-1.

The records are generated in a 10-day simulation with a time step of 15 minutes (this is time step considered by the water board as useful for the management of the area and the pump operation). Additionally, the parameter $a$ in Equation (5-1) used in the quantization of the time series is determined by looking at the water level variations that are negligible for management. In our case, $a=5$ (cm) is the minimum dimensional unit for which the water management becomes critical. For example, water level variations of less than 5cm can be due to wind, ship movement or dynamic waves generated by the operation of pumping stations and are not important for water management (that is such variations should not require the hydraulic structures to be operated) and so they should be considered as noise in the time series. (i.e., these time series should not provide information content, though they would have high entropy values without discretization). As an example, the water level time series at the discharge of one of the pump stations is presented in Figure 5-2, together with its quantized version, used in the entropy calculation. Finally, the value of $\epsilon$ in step g) is considered to be the mean of the vectors $v_i$, so that half of the points were considered independent and half dependent. This assumption is valid since only points with low pairwise dependency criteria with respect to a given point are selected.

5.2.3 Analysis of Results

In order to have an initial idea of the behaviour of the information content of the Pijnacker region polder system, the maps of marginal entropy and mutual information are drawn. In the first place the entropy map of the system is shown in Figure 5-3 where several static information zones can be identified. A few points with null entropy can be found, which correspond to fixed boundary conditions for water level (that is, pump stations discharging to big storage bodies with fixed water level, to the southeast of the
area). This map provides a first view of the areas (with high entropy) where it is suitable to place the first water level monitoring station, from the information content viewpoint.

![Quantized water level and original water level](image)

Figure 5-2. Example of original water level time-series and its quantized version, at a point located downstream a pumping station.

![Entropy map of Pijnacker region](image)

Figure 5-3. Entropy map of Pijnacker region
The subsequent monitors must be as independent as possible to this first monitor. In the second place, in order to highlight the information-dependency in the system, Figure 5-4 shows the \( DIT \) index calculated between an arbitrary point \( A \) and the rest of the system. In the figure, the darker the point, the greater is the dependency of points with the point \( A \). The dependency of \( A \) on its neighbouring points is evident. However, some other regions seem to have a strong dependency with \( A \) in spite of their hydraulic disconnection. It can also be noticed that only a few points (the ones with constant water level during the time of simulation) are completely independent (\( DIT=0 \)).

\[
\text{Figure 5-4. Directional Information Transfer Index (DIT) map for the point A (bits).}
\]

In order to look for a set of \( N \) gauging stations that can provide the maximum information of water levels in the system, the WMP approach is used, having the dependency criteria \( I(X;Y) \), \( DIT_{XY} \), and \( DIT_{YX} \) as the basis to create the matrix \( T \) of Equation (5-4). In order to facilitate further analyses, \( n \) is initially considered to be 9. For the first criterion \( I(X;Y) \), Figure 5-5 is obtained, where, beside the monitor locations, the overlapped \( I(X;Y) \) map at each step and the value of the marginal entropy of the placed monitor is presented. The darker the calculation point, the higher its dependency with respect to the set of previously selected monitors.

Similarly, Figure 5-6 is obtained applying the second criterion \( DIT_{XY} \) in the WMP approach. In this case, a value of 1 was assigned to the dependent points and 0 to the independent ones, in such a way that the empty areas indicate where the next monitor must be placed. It can be noticed that the ninth monitor does not have independent points.
associated to it, implying that this ninth monitor is not needed. This is also confirmed by its small value of entropy $H(X_9) = 1.38$ bits, which does not provide much more information to the joint set.

![Image](image.png)

Figure 5-5. Step-by-step solution for the location of water level monitors using $I(X;Y)$ as the dependency criteria. Entropy of the currently selected point is shown at each step.

Finally, WMP is performed using the third dependency criterion $DIT_{XY}$, obtaining the nine monitors presented in Figure 5-7.

A small area to the south with no coverage can be noted, which corresponds to points with very low entropy values (see Figure 5-3). This implies that any point in the system is completely independent of any point that belongs to this area (Figure 5-4, for example, confirms this statement for the case of point $A$). This situation explains why the ninth monitor in the previous experiment is not worth being selected.

The location of the sets of monitors obtained by means of the three criteria is shown in detail in Figure 5-8, where the sequences of monitor selection are not presented for clarity. The summary of the monitors obtained with the different independency criteria is presented in Table 5-1, as well as their corresponding values of Total Correlation and Joint Entropy, the latter being calculated from Equation (2-9).
Optimisation of monitoring networks for water systems

Figure 5-6. Step-by-step solution for the location of water level monitors using $DIT_{xy}$. Entropy of the currently selected point is shown at each step (bits).

Figure 5-7. Step-by-step solution for the location of water level monitors using $DIT_{yx}$. 
Figure 5-8. Location of water level monitors obtained by the WMP approach, using \( I(X;Y) \), \( DIT_{XY} \) and \( DIT_{XY} \) as pairwise dependency criteria.

5.2.4 Discussion

The solutions obtained with the WMP method are discussed and evaluated in this section. In the first instance, the three pairwise criteria give some similar monitor locations in terms of spatial distribution. Besides the monitor at point 441 (selected by the three solutions, since the approach starts with the point with highest entropy), it is noticed that the monitors at points 733, 133, 426 and 438 have been selected by at least two of the solutions. Besides this, there are points that are separately identified but are the same from the practical point of view, such as points 319 and 320, 426 and 876, 719 and 88 as well as 133 and 144.

It can be noticed, for each criteria of dependency used, that every time a monitor is added to the set, the number of independent points is reduced, and that this reduction becomes less evident when new monitors are added. A quantitative way of looking at the reduction of uncertainty in the system when a new monitor is selected is by evaluating the value of joint entropy at every step. Since the calculation points are not completely independent, then \( H(X_1, X_2, \ldots, X_N) \neq \sum_{i=1}^{N} H(X_i) \), so the concept of Total Correlation is needed to evaluate the multivariate independency, by subtracting the summation of the marginal entropies from the value of Total Correlation estimated using the grouping property. Figure 5-9 shows that the three solutions have a similar behaviour for both information content and independency. Furthermore, the reduction in uncertainty is strong for the first
monitors, due to the fact that more information among them is shared as new monitors come into the solution. This explains why additional points do not reduce the uncertainty in the system.

In order to allow further analysis, the following values are computed:

\[
C(X_1, X_2, \ldots, X_{n=1520}) = 3510 \text{ bits}
\]

\[
\sum_{i=1}^{n=1520} H(X_i) = 3519.2 \text{ bits}
\]

\[
H(X_1, X_2, \ldots, X_{n=1520}) = \sum_{i=1}^{n=1520} H(X_i) - C(X_1, X_2, \ldots, X_{n=1520}) = 9.2 \text{ bits}
\]

It is clear that the total correlation of all points in the system almost equals the sum of their marginal entropies. This means that the amount of information shared between all of the points in the system is practically the same as the total amount of information that each point adds to the system separately. Moreover, the maximum amount of information content that can be extracted from the system is 9.2 bits.

The implication of having \( C \) almost as big as the sum of the marginal entropies is that the calculation points are highly dependent on each other. This is like having a Venn diagram with 1520 circles of different size that overlap each other almost completely. If all the
circles were independent from each other, then $C=0$ and they do not overlap. The problem could be seen as selecting the best $N$ circles that “cover” the total “area” of the $n$ circles but that at the same time have little “overlapping area”.

Table 5-1. Summary of monitors obtained by each dependency criteria and corresponding values of joint entropy and total correlation

<table>
<thead>
<tr>
<th>Criteria</th>
<th>ID of selected monitor (in order)</th>
<th>Joint Entropy $H(X_1, X_2, \ldots, X_9)$</th>
<th>Total Correlation $C(X_1, X_2, \ldots, X_9)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I(X,Y)$</td>
<td>441 133 320 948 876 477 72 272 795</td>
<td>7.32</td>
<td>16.79</td>
</tr>
<tr>
<td>$DIT_{XY}$</td>
<td>441 254 446 905 144 313 719 733 438</td>
<td>7.63</td>
<td>17.7</td>
</tr>
<tr>
<td>$DIT_{YX}$</td>
<td>441 133 426 319 88 730 29 733 673</td>
<td>6.82</td>
<td>13.88</td>
</tr>
</tbody>
</table>

The results are summarized in Table 5-1, in which the value of the total joint entropy and the Total Correlation are included for each solution. From Figure 5-9 it can be observed that with only the first two monitors is possible to reach 5.5 bits of information content (60% of the total information content of the system), at a relatively low dependency value (less than 0.5% of the total correlation of the system). This provides a good criterion for evaluating the quality of the results.

In general, the monitors selected are located next to hydraulic structures. This can be explained by the fact that these elements provide high, systematic variations to the water levels in the system, during a precipitation event. Although entropy increases considerably by water level variations from pump stations, its informative capability is not important if such variations occur within a small water level range. In this case, quantization appears to be a promising approach to filter out these noisy signals.

The parameter $a$ of Equation (5-1) can be viewed as the minimum dimensional unit of water level for which the management of the system becomes critical. In the case of a typical low-lying polder system in the Netherlands, 5cm is already decisive for water management in terms of the operation of control structures. Conversely, water level variations smaller than 5cm (due, say, to wind, ship movement or dynamic waves generated by the operation of pumping stations), are not important for water management and so they should be considered as noise in the time series. In other water systems such as rivers, $a=5$cm might be too low.

A sensitivity analysis of $a$ is carried out by comparing monitor locations obtained for integer values of $a$ between 1 and 10, and for 15 and 20 cm. The comparison shows that for each value of $a$ between 3cm and 8cm a minimum of 55% of the same locations are shared with the average of the four neighbouring solutions that are the closest in number. For values of $a < 3$ cm and $a > 10$ cm, only a few monitor locations are common. Figure 5-10 shows the average percentage of common monitors when comparing solutions for different values of $a$. In general, 25% of the locations are common to all solutions, although they are not necessarily selected in the same order. The first monitor location is
common to all solutions (which is the one with the highest entropy value, as explained in step d).

![Percentage of common monitor locations](image)

Figure 5-10. Average percentage of common monitor locations comparing the solution obtained for each value of \( a \) with all other calculated solutions (\( a=1,2,\ldots,10,15,20 \))

From the hydraulics point-of-view, it is important to note that the monitors located downstream of a weir cannot give any information about the conditions upstream of the weir when this is working in a modular regime. On the other hand, when the weir regime is drowned, the downstream monitoring point may provide additional information to the system upstream. In general terms, all the weirs break the dependency in information between upstream and downstream points; in a similar way they break the continuity in water levels. However, some weirs working in the modular regime may not cause a discontinuity in information because the same local hydrological information may be shared upstream and downstream. This situation can be seen in Figure 5-3 and Figure 5-4, in which not all the weirs limit the information areas.

The gauge locations will not change significantly when using different time steps in the hydrodynamic model, as long as it acceptably replicates the real behaviour of the system. Certainly, the quantization of the time series as well as the frequency-based method to estimate the probabilities for the information-related measures would give similar entropy values. Additionally, the relative nature of the pairwise criteria makes entropy values unchanged using different time steps.

Even though values of \( \varepsilon \) equal to 0 will define exactly the set of points that is dependant on a particular point, its applicability might lead to an empty set of independent values, because even a small amount of information may be shared between hydraulically disconnected points. In this case the points can be hydrologically-related; for example, a precipitation event that affects water levels at two hydraulically-unrelated points may
Chapter 5 - Information Theory for monitor location

induce a correlation between them and thus share some of their information content, and \( I(X_i; X_j) > 0 \). This is the reason why the independent points were selected considering the mean of the independency criteria as a threshold.

The fact that the monitored signals are dependent is, in practice, a desirable condition, since cross-checking is necessary to validate the data from other stations and to detect possible errors. Nevertheless, the theoretical design of monitoring networks should still consider the most independent monitors as priority places for data collection.

5.2.5 Conclusions

A number of “information zones” can be identified in both entropy and transinformation maps, defined by pump stations and by weirs working in modular regime, which create discontinuities in the information content upstream and downstream of the structures. One information zone may include several target water levels.

The solutions obtained with the three pairwise criteria \( I, DIT_{XY} \) and \( DIT_{YX} \) have some monitors in common in terms of spatial distribution. However, the solution obtained with the \( DIT_{XY} \) criterion provides the highest value of joint entropy for the set of monitors and the highest value of Total Correlation. A high value of joint entropy reveals a high information content. This is the measure that identifies the preferred solution for the monitors.

On the contrary, the solution obtained with the \( DIT_{YX} \) criterion provides the lowest information content and the minimum value of Total Correlation. According to the conditions mentioned in the procedure for gauge location, this implies that \( DIT_{YX} \) is more effective to fulfil condition a) (independency), whereas \( DIT_{XY} \) is better to fulfil condition b) (amount of information content). This opens a new possibility to solve the problem of monitor selection out of a dense set of potential monitors, which is to use multiobjective optimization where \( DIT_{YX} \) is to be maximized and \( DIT_{YX} \) minimized.

Results show that the calculation points may be highly dependent on each other even if some of them are hydraulically disconnected. This dependency is due to the hydrological connection between the points, since in relative small areas the same rainfall events are shared. Although these hydrological dependencies make the problem of looking for independent monitors more difficult, the proposed methodology proves to be a suitable method for this purpose.

The values of marginal entropy are sensitive for different values of \( a \) used in Eq.(5-1). Small values lead to high entropy values for locations near pumping stations, whereas bigger values tend to filter out small disturbances, causing a decrease in entropy. However, as \( a \) affects all the generated time series equally, the selection of the first monitor location (see step d in the section Location of Gauges) remains the same for different values of \( a \). On the contrary, transinformation and \( DIT \) values do not change at all, due to the relative nature of the expressions. The value of \( a=5 \) (cm), however, is found to be the minimum dimensional unit for which the water management becomes
critical and seems to be good for keeping the informational property from the management point of view.
The use of multiobjective optimization techniques to solve the problem is explored in the next section, using the minimization of the total correlation as a first objective and the maximization of the joint entropy as a second objective; in addition, practical considerations as constraints are also included.

5.3 Optimising Information measures for the design and evaluation of monitoring networks in polders

A method for siting water level monitors based on information-theory measurements is presented. The first measurement is Joint Entropy, which evaluates the amount of information content that a monitoring set is able to collect, and the second measurement is Total Correlation, which evaluates the level of dependency or redundancy among monitors in the set. In order to find the most convenient set of places to put monitors from a large number of potential sites, a Multi Objective Optimization Problem (MOOP) is posed under two different considerations: 1) taking into account the costs of placing new monitors, and 2) considering the cost of placing monitors too close to hydraulic structures. In both cases, the joint entropy of the set is maximized and its total correlation is minimized. The costs are considered in terms of information theory units, for which additional terms affecting the objective functions are introduced. The proposed method is applied in a case study of Delfland region, The Netherlands. Results show that Total Correlation is an effective way to measure multivariate independency, and that it must be combined with Joint Entropy to get results that cover a significant proportion of the total information content of the system.

5.3.1 Optimising Joint Entropy and Total Correlation: justification

Regarding the use of Information Theory in the design of monitoring networks, the analysis is based on looking for those locations where the information content about a particular water-related variable is a maximum, so that a monitor device placed there has “potential information” in the sense that once placed, it would reduce uncertainty by providing information (Mogheir and Singh 2002). For more than one variable (that is, more than one monitoring device), Joint Entropy is used, because it represents the information content of the set of monitoring devices, which can be maximized, as Caselton and Husain (1980) did for the case of reducing an existing rainfall monitoring network.

Additionally, minimizing transinformation between monitors is the basis of designing monitoring networks by applying Information Theory. Naturally, the placement of two monitors that provide exactly the same information is not optimum. In other words, the redundancy of the monitors should be as small as possible (Mishra and Coulibaly 2009). In this Chapter, Total Correlation is used to measure redundancy among multiple variables.
The main contribution of this method is that Joint Entropy and Total Correlation are independent objectives that must be optimized. A Venn diagram can be used to illustrate this idea, where the information content of a monitor is represented by the area of a circle; the information shared between variables corresponds to their overlapping areas and the information content of the set of variables is described by the total area covered by the circles (see e.g., Cover and Thomas 1991; Ruddell and Kumar 2009). Suppose, for example, that ten potential locations are available to place three monitors; the equivalent Venn diagram might appear as in Figure 5-11(A) which represents the information content of 10 variables and their common information.

![Venn Diagram](image)

Figure 5-11. Venn diagrams illustrating the proposed optimization problem. (A): Information content of 10 variables and their common information; (B), (C) and (D): possible solutions for the selection of three monitor locations (1), obtained by maximizing joint entropy (2) and minimizing Total Correlation (3).

The task is to select the most informative set of three variables that, simultaneously, are least interdependent, i.e., the summation of the overlapped areas is a minimum. Three possible solutions for this generic example are shown in Figure 5-11(B), (C) and (D). The relationships in terms of information content for each solution are shown in the first row of Figure 5-11; the area representing the Joint Entropy of each solution set is shown by the total covered area in the second row of Figure 5-11 (to be maximized) and the overlapping areas representing Total Correlation is shown in the third row (to be minimized).
5.3.2 Description of the MOOP methodology

The Multi Objective Optimization Problem (MOOP) is posed to find the set of new stations \( X = \{X_1, X_2, \ldots, X_M\} \) that optimally complement the set of \( N \) already existing stations \( E = \{E_1, E_2, \ldots, E_N\} \), in such a way that the joint set of \( M+N \) stations \( S = \{E, X\} \) provides the maximum possible information content with the minimum shared information between them. The first objective is described by the joint entropy, Eq. (2-4), while the second is described by the Total Correlation, which is estimated following the steps described in the previous section. To understand why the second objective is needed the situation in which two water level gauges are located extremely close to each other can be considered; they would in effect record the same time series, they would have the same information content and therefore both stations would be completely dependent (redundant).

Additionally, every single station should be placed where the highest information content can be extracted. Under this consideration, water system points with constant or quasi-constant water-level records should not be selected because they do not provide any further information (i.e., water level value does not change, so just one record would be enough to describe the state of that point). This consideration can be added as a constraint in the MOOP by excluding low-entropy points from the decision variable set, which in turn implies a reduction in the search space and therefore a reduction of the computational effort. The definition of the threshold to identify low-entropy points is described in the next section.

Taking into account the previous considerations, the multi-objective optimization problem (MOOP) is mathematically formulated as follows, Eq (5-5).

\[
\begin{align*}
\text{min} \left\{ C(S_{M,N}) = C(X_1, X_2, \ldots, X_M, E_1, E_2, \ldots, E_N) \right\} \\
\text{max} \left\{ H(S_{M,N}) = H(X_1, X_2, \ldots, X_M, E_1, E_2, \ldots, E_N) \right\}
\end{align*}
\]

subject to:

\[ H(X_i) > \text{Low entropy threshold} \]
\[ N + M = \text{number of monitors} \]

(5-5)

The threshold to define low-entropy points to be discarded from the search space was defined by looking at the relative frequency of the entropies in the system. The Figure 5-12 shows that more than 50% of the points have an entropy value below 0.1 bits, which represents less than 7% of the point with maximum entropy. This implies that the search space is dramatically reduced by a factor of several hundreds. This value is used in Eq. (5-6) and Eq. (5-7) introduced below.

The procedure to estimate the probabilities required for the calculation of Eq. (2-1), (2-4) and (2-8) follows a frequency analysis described in the section 5.2. Moreover, the MOOP is solved using the Non-Sorted Genetic Algorithm (NSGA-II) by Deb et al. (2002) and implemented in the NSGAX software (Barreto et al. 2006).
5.3.3 Case study: Pijnacker region, The Netherlands

The developed methods were applied in the same case study used in section 5.2, a polder system in a sub-district of Delfland, in The Netherlands (see details in Chapter 3). The existing water-level monitoring network fails to provide the proper information about the state of the system, especially under extreme weather conditions. This is because the measurements are basically used to check whether the current water levels are between the on/off levels of the pumps and to operate them accordingly. The measurement points of the 9 biggest pumping stations are further identified as Existing Pump Monitors (EPM). The hydrodynamic model used to generate the water level time series was run with a typical rainfall event for the area. For this study a rainfall event of 5-years return period was used (much less extreme than the one used in section 5.2, as it is of particular interest for the management of the region in the assessment of regular flood risks.

5.3.4 Practical situations included in the optimization problem

Two practical situations occur in the Delfland water system: a) the need to introduce financial restrictions on installing new monitors, and b) the problem of the accuracy of measurements taken near hydraulic structures, due to small water level fluctuations. Although both situations should be considered simultaneously, they are applied separately in order to facilitate the analysis of the results. For both situations it is assumed that 9 monitors in total need to be located in the system.

- Approach a)

For the first situation we define an additional term $u*M$, representing the cost (in informative units) of having to place $M$ new monitors. The parameter $u$ is a constant with cost units of bits per new monitor. The term $u*M$ is then added to the total correlation ($u$
Optimisation of monitoring networks for water systems

bits of redundancy are added to the set) and subtracted from the joint entropy ($\mu$ bits of joint information are subtracted from the set) every time the set contains a new monitor. In this way, the optima are kept separate for both objectives as $M$ increases. Although $\mu$ may differ according to the location of a particular monitor, a constant value is considered to simplify the problem. For the subsequent experiments, $\mu$ is defined equal to 1 bit/new monitor; a sensitivity analysis of this parameter is presented at the end of this section. The resultant optimization problem can be written as:

$$\begin{align*}
\min \left\{ F_1 = C(S_{M,N}) + \mu M = C \left( X_1, X_2, \ldots, X_M, E_1, E_2, \ldots, E_N \right) + \mu M \right\} \\
\max \left\{ F_2 = H \left( S_{M,N} \right) - \mu M = H \left( X_1, X_2, \ldots, X_M, E_1, E_2, \ldots, E_N \right) - \mu M \right\}
\end{align*}$$

subject to,

$$H \left( X_i \right) > 0.1 \text{bits}$$

$$0 \leq M \leq 9; 0 \leq N \leq 9; M + N = 9$$

Approach b)

For the second situation we introduce the term $q^*v$, which represents the cost in informative units, of having to place monitors close to hydraulic structures. $v$ is defined as the number of times the distance $d_s$ (minimum distance between a monitor and a structure evaluated over all possible combinations of monitors and structures) is violated by a particular solution; $q$ is a constant with cost units of bits per violation of minimum distance. It is assumed $d_s=50m$ and $q=1$ bit/violated $d_s$. As in the first situation, the term $q^*v$ is added to the total correlation and subtracted from the joint entropy of each evaluated set in order to keep the optima away from the ideal point (min $C$ and max $H$) as $v$ increases. The resultant optimization problem can be written as shown in Eq. (5-7)

$$\begin{align*}
\min \left\{ F_1 = C(S_{M,N}) + qv = C \left( X_1, X_2, \ldots, X_M, E_1, E_2, \ldots, E_N \right) + qv \right\} \\
\max \left\{ F_2 = H \left( S_{M,N} \right) - qv = H \left( X_1, X_2, \ldots, X_M, E_1, E_2, \ldots, E_N \right) - qv \right\}
\end{align*}$$

subject to,

$$H \left( X_i \right) > 0.1 \text{bits}$$

$$0 \leq M \leq 9; 0 \leq N \leq 9; M + N = 9$$

In order to solve Eq. (5-6) and Eq. (5-7), the Non-Sorted Genetic Algorithm (NSGA-II), (Deb et al. 2002), was used with the following evolutionary parameters: crossover probability = 90%, mutation probability = 1/9. The evolutionary parameters of number of population and generations were set after several experiments with different values; the use of 500 populations and 2000 generations were found convenient, because with bigger values the solutions do not improve significantly for the two situations.

5.3.5 Analysis of Results

In order to facilitate the analysis, the calculation points of the system have been labelled with integer numbers from 1 to 1520. It is worth noting that the total joint entropy of the system (i.e., the joint information contained in these 1520 points as a single set) is
Chapter 5 - Information Theory for monitor location

\[ H_{sys} = H(X_1, X_2, ..., X_{1520}) = 4.91 \text{ bits}, \] a value that represents the ideal amount of information that the network of monitors \( S \) should provide. It must be noted that in previous approach (section 5.2) the value for \( H_{sys} = 9.2 \text{ bits} \) was obtained using a rainfall event with a much higher return period. The results obtained are evaluated with respect to this value.

- Results for approach a)

The Pareto-optimal set of solutions obtained for the first situation is presented in Figure 5-13, where the solutions are characterized according to the number of existing monitors that were picked up in the optimization process (0, 1, 2 and 3 EPM). For this point onwards, the notation \( S_{M,N} \) will be used to show that the set of monitors \( S \) is composed of \( M \) new monitors and \( N \) existing monitors.

![Figure 5-13 Pareto-optimal set of solutions discriminated by EPM, approach a). Extremes \( X_a \) and \( Y_a \) are indicated for further analysis. Results obtained with WMP method (Alfonso et al. 2010b) are also indicated.](image)

The solution \( S_{0,9} \) (corresponding to the full set of EPM) is not presented in the figure because it is out of the figure scale, but its location can be observed in Figure 5-14. This solution has a very small total correlation (close to 0) but also a relatively small information content (1.04 bits): one record taken at each of these monitors would jointly provide slightly more than one bit of information on average, or about 20% of the information of the state of the system, meaning that the current monitoring network is far from optimal.
Several interesting facts can be mentioned (Figure 5-13):

- Solutions for \( N = 4, 5, 6, 7, 8 \) and 9 are always dominated by other solutions, so they are not found to be part of the Pareto front.
- Existing monitors 541, 1337, 669, 842 are not selected in any scenario.
- Solutions for \( S_{3,6} \) always include monitors 1187, 465 and 56. These \( N = 6 \) (new) monitors make the joint entropy vary between 3 and 3.5 bits (between 60% and 70% of \( H_{sys} \)) and between 0.2 and 0.6 bits in total correlation terms.
- All the previously commented solutions (\( S_{3,6} \)) are always dominated by solutions that consider fewer existing monitors and more new monitors in the final set.
- For the scenario \( S_{2,7} \) we found solutions with joint entropy between 3 and 4 bits, that range between 0.2 and 1.1 bits of total correlation. Only three combinations of two existing monitors (1187, 56); (1187, 465) and (1203, 56) are part of the Pareto front of optimal solutions.
- From Figure 5-13 it is clear that this Pareto front is closer to the ideal value (\( C = 0, H_{J} = 4.91 \)) and therefore it dominates the previously discussed solution \( S_{3,6} \).
For the case of $S_{1,8}$, the resultant Pareto front dominates practically all the solutions obtained for the previously discussed scenarios.

In order to characterize the solutions, the extremes of the Pareto front are analyzed (Figure 5-13), where $X_a$ identifies the solution that maximizes the Joint Entropy (bottom-right extreme) and $Y_a$ identifies the solution that minimizes the Total Correlation (upper-left extreme). The sub-index $a$ is provided to distinguish the solutions obtained for approach a and approach b. First, the solution at $X_a$, which maximizes the (negative) joint entropy, places (also new) monitors at $S_{9,0} = (587, 991, 286, 1030, 57, 458, 42, 175, 1204)$, with joint entropy of 4.18 bits (85% of $H_{sys}$) and total correlation of 1.19 bits. Second, the solution at $Y_a$, which minimizes the total correlation, corresponds to the selection of (all new) monitors $S_{9,0} = (827, 1490, 1078, 1265, 620, 704, 891, 394, 1151)$, which is a set with total correlation of 0.0 and joint entropy of 1.51 bits (30% of $H_{sys}$).

These two sets of monitors $X_a$ and $Y_a$, as well as the solution for which $S=EPM$, are located in the map of the system in Figure 5-14. As expected, the monitoring sets are different in spatial terms. The most important monitor of each extreme, in terms of information, are the monitors $X_{a6}$ (point 458) and $Y_{a7}$ (point 891), because they are located in a zone with high marginal entropy. However, in both extremes $X_a$ and $Y_a$ (and regardless of having excluded the points with the lowest entropy) points with low entropy are included in the solutions. On one hand, at the extreme $X_a$, the point $X_{a2}=991$ has a low information content, which means that the eight remaining points would be enough to place. This situation may lead to a refined criterion to determine the number of monitors to be placed, so that no assumptions in this regard would be needed and the second constraint of equations (5-6) and (5-7) may not be considered.

On the other hand, all the monitors of the solution at extreme $Y_a$ are located at very low informative sites, with exception of point $Y_{a6} = 704$ with $H(Y_{a6})=1.5$ bits. This explains why this solution has such a low Total Correlation: low-entropy points are more independent from the rest than high-entropy points. Naturally, this solution is far from being a good set for monitoring because it does not provide significant joint information.

- Results of approach b)

Following the same identification pattern used for the approach a), the extremes of the Pareto front in Figure 5-15, is used, being $X_b$ the solution that maximizes the Joint Entropy and $Y_b$ the solution that minimizes the Total Correlation. Additionally, Figure 5-15 discriminates the solutions by the number of violations $v$ of the minimum distance $d_s$. Several observations can be made. First, it can be noticed that, in spite of having only 9 monitors to place, some of the solutions have violated the distance rule 10 times, which means that one monitor was close to either a weir or a pump station more than once. This was expected because of the high density of hydraulic structures in the area. Second, solutions with low total correlation are found only when no violations take place. This implies that non redundant sets of monitors are only possible to place away from hydraulic structures. However, the price of this independency is paid by the fact that jointly they collect relatively low information (less than 60% of the information content of the system); the trade-off between the two information quantities is again evident. In
third place, three solutions give the highest joint entropy, and correspond to solutions with 3, 4 and 5 distance violations.

A detailed analysis of the optimization is presented in Table 5-2, which categorizes the 500 solutions obtained for approach b) in terms of violations of the minimum distance due to the presence of pumps and/or weirs. It can be noted that the majority of violations are caused by pumps for solutions that include combinations of violations (by pumps and by weirs). For instance, for 6 violations, the combined possibilities are 5 pumps + 1 weir (3 solutions), 4 pumps + 2 weirs (22 solutions) and 3 pumps + 3 weirs (18 solutions), Table 5-2. In other words, the number of violations due to the proximity of the monitors to the pumps is generally bigger than the number of violations due to their proximity to the weirs. This can be explained by the fact that a pump operation adds entropy to the upstream and downstream neighbouring points while a fixed weir in modular regime stabilizes the water levels in a way that downstream points reduce their marginal entropy. However, no trivial pattern in the Pareto front was identified based on the number of pumps or the combination of pumps and weirs.

The resulting monitoring sets obtained at both extremes of the Pareto front (for maximum joint entropy and for minimum total correlation) are located in the map of the water system in Figure 5-16. The solution for the extreme Xb, gives a maximum joint entropy of
Chapter 5 - Information Theory for monitor location

4.04 bits, about 82% of $H_{sys}$. However, two points appear to be low-informative; these are $X_{b2}$ (point 776) and $X_{b9}$ (point 170). Similar to the first situation, it has been found that nine monitors are not necessary: indeed seven are enough to describe the information content of the system. For the case of the extreme $Y_b$, we find again a similar situation as in the first situation: the majority of the computational points have negligible information content, with the exception of points $Y_{b3}$ (point 1030) and $Y_{b7}$ (point 801). This result shows again that points with very low entropy contribute to decrease the total correlation of the set. However, as in the previous set, this solution should not be taken into account because it provides low information content (about 2.0 bits or 40% of $H_{sys}$).

Table 5-2. Number of solutions for approach b) with minimum distance violations by pumps and weirs.

<table>
<thead>
<tr>
<th>Number of solutions with violations by pumps</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>79</td>
<td>21</td>
<td>8</td>
<td>6</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>118</td>
</tr>
<tr>
<td>1</td>
<td>30</td>
<td>33</td>
<td>30</td>
<td>30</td>
<td>18</td>
<td>3</td>
<td>-</td>
<td>144</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>16</td>
<td>15</td>
<td>19</td>
<td>22</td>
<td>23</td>
<td>-</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>-</td>
<td>14</td>
<td>18</td>
<td>26</td>
<td>31</td>
<td>7</td>
<td>96</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>8</td>
<td>19</td>
<td>15</td>
<td>42</td>
</tr>
<tr>
<td>Total</td>
<td>114</td>
<td>70</td>
<td>67</td>
<td>73</td>
<td>78</td>
<td>76</td>
<td>22</td>
<td>500</td>
</tr>
</tbody>
</table>

- Comparison of results with WMP approach
The set of monitoring networks obtained through the MOOP for both approaches a) and b) are compared with the method Water Level Monitoring Design in Polders, WMP (Alfonso et al. 2010b), in which three pairwise criteria, namely Transinformation $I(X,Y)$, and Directional Information Transfer, $DIT_{XY}$ and $DITY_{X}$ (Yang and Burn 1994) are used to evaluate the independency of the monitoring set. The WMP is a step-based method that can be classified as a greedy algorithm, in which the next best monitor (with high information content and low dependency) is selected at each step, given the total number of monitors. For the sake of comparing results, the WMP method was run for the same rainfall event used in this paper. It must be noted that the WMP method was run with no constraints, so the resultant monitoring network is composed of only new devices, and their proximity to hydraulic structures is not taken into account.

The information theory characteristics of the monitoring sets obtained with the WMP method are included in the Pareto-optimal solutions of Figure 5-13 and Figure 5-15. The result for $I(X;Y)$ is almost identical to the result for $DIT_{XX}$ so they are presented as a single point in the graphs and the result for $DIT_{XY}$ is out of the scale of both graphs. The immediate conclusion is that the WMP method provides solutions that are part of both the Pareto-optimal front obtained for sets composed by new monitors only (Figure 5-13) and the Pareto-optimal front obtained for sets that do not violate the minimum distance to hydraulic structures.
5.3.6 Discussion

5.3.6.1 Priority of monitors’ placement

One of the characteristics of the multiobjective optimization with genetic algorithms is that during each step of the process, all the variables (monitors) that belong to each solution are generated at the same time, regardless of their individual significance in informative terms. Nevertheless, this is an important issue when placing the monitors, because it is expected that during their implementation some monitors have a different priority than others. In order to prioritize the monitors, again from the information standpoint, it is possible to sort them either by total correlation (in ascending order, so the monitor that adds the least $C$ to the set is placed first) or by joint entropy (in descending order, so the monitor that adds the biggest $JH$ to the set is placed first).

5.3.6.2 Approach a)

Figure 5-17 shows the behaviour of the information theory values for the solutions of approach a) obtained at the extremes $X_a$ (left) and $Y_a$ (right), sorting them by total correlation in ascending order (top), and by joint entropy in descending order (bottom).
Chapter 5 - Information Theory for monitor location

5.3.6.3 Approach b)

Figure 5-18 presents the previous analysis for the results obtained for the second situation. In this case it is evident that monitors 776 and 170 obtained in the solution for the extreme X_b do not add any information to them jointly, \( H(1229)=H(1229,776,170) \) so they have no capacity to reduce the total correlation to the set. As in the first situation, the extreme for which total correlation is a minimum (extreme Y_b), shows that only the
points 1030 and 801 are informative points, \( H(1030,801) \), and therefore only 7 points would be needed. This confirms again that the solutions located at this extreme of the Pareto front should not be considered for the final selected monitoring set.

![Figure 5-18. Progress of information quantities as new monitors are added for solution of approach a), for extremes \( X_a \) and \( Y_b \) of Figure 5-15.](image)

5.3.6.4 Sensitivity analysis of the parameters \( u \) and \( q \)

The initial assumptions of \( u=1 \) and \( q=1 \) were initially chosen to express costs in informative units, and their selection may affect the outcomes. In order to make the sensitivity analysis, \( Eq (5-6) \) was solved for values of \( u \) of 0.1, 0.5, 1, 2 and 5 bits/new monitor. In order to evaluate the quality of the results, the extremes of the obtained Pareto fronts are analyzed. Additionally, the average of the marginal entropy of each solution is analyzed, in order to assess the distribution of the entropy among the selected monitors.

![Figure 5-19. Sensitivity of the maximum Joint Entropy (1) and Total Correlation (2) due to variations of the parameter \( u \), discriminated by the number of new monitors in the solution.](image)
Chapter 5 - Information Theory for monitor location

Figure 5-19 and Table 5-3 have been prepared for the parameter $u$, where the sensitivity of the maximum Joint Entropy and the Total Correlation due to variations of the parameter $u$, is presented, according to the number of new monitors in the solution. It can be observed that, for solutions having only new monitors, good results are achieved regardless of the value of $u$ (Joint Entropy between 4.12 and 4.38 bits or between 84% and 89% of $H_{sys}$).

However, solutions considering some existing monitors become less informative as $u$ decreases. This is because there are no restrictions on placing the new monitors and better combinations of monitors are found. An additional effect is that no solution that considers 4 EMPs was found within the solutions for $u=0.1, 0.5$ and $1.0$. Conversely, high values of $u$ force the inclusion of existing monitors, which generates solutions of lower quality (lower information content; see Figure 5-19(1)). When looking at the distribution of the Total Correlation (Figure 5-19-(2)), solutions with independent monitors are obtained when more existing monitors are considered, situation that again supports the conflicting nature of the objectives.

Table 5-3. Sensitivity analysis for parameter $u$.

<table>
<thead>
<tr>
<th>Criteria for evaluation of solutions</th>
<th>EPM</th>
<th>0.1</th>
<th>0.5</th>
<th>1</th>
<th>2</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Joint Entropy obtained</td>
<td>0</td>
<td>4.12</td>
<td>4.26</td>
<td>4.31</td>
<td>4.27</td>
<td>4.38</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3.52</td>
<td>3.84</td>
<td>3.73</td>
<td>4.20</td>
<td>4.37</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3.04</td>
<td>3.54</td>
<td>2.95</td>
<td>4.12</td>
<td>4.30</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2.75</td>
<td>2.85</td>
<td>2.10</td>
<td>3.71</td>
<td>4.18</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3.36</td>
<td>2.61</td>
</tr>
<tr>
<td>Min Total Correlation obtained</td>
<td>0</td>
<td>1.27</td>
<td>2.40</td>
<td>1.91</td>
<td>1.60</td>
<td>2.57</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.47</td>
<td>0.85</td>
<td>1.06</td>
<td>1.87</td>
<td>2.10</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.19</td>
<td>0.58</td>
<td>0.89</td>
<td>1.88</td>
<td>2.54</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.13</td>
<td>0.14</td>
<td>0.07</td>
<td>1.43</td>
<td>1.88</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.34</td>
<td>0.52</td>
</tr>
</tbody>
</table>

Figure 5-20. Sensitivity of the maximum Joint Entropy (left) and Total Correlation (right) due to variations of the parameter $q$, discriminated by the number of new monitors in the solution.
A similar analysis was carried out for the parameter $q$, see Figure 5-20 and Table 5-4. In general, solutions obtained with low values of $q$ seem to be less informative and less independent, regardless of the number of violations of distance to hydraulic structures. Additionally, a value of $q = 0.1$ gives solutions that do not include a high number of violations (see Table 5-4).

Table 5-4. Sensitivity analysis for parameter $q$.

<table>
<thead>
<tr>
<th>Criteria for evaluation of solutions</th>
<th>EPM</th>
<th>0.1</th>
<th>0.5</th>
<th>1</th>
<th>2</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Joint Entropy obtained</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>4.12</td>
<td>4.26</td>
<td>4.31</td>
<td>4.27</td>
<td>4.38</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3.52</td>
<td>3.84</td>
<td>3.73</td>
<td>4.20</td>
<td>4.37</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3.04</td>
<td>3.54</td>
<td>2.95</td>
<td>4.12</td>
<td>4.30</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2.75</td>
<td>2.85</td>
<td>2.10</td>
<td>3.71</td>
<td>4.18</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3.36</td>
<td>2.61</td>
<td></td>
</tr>
<tr>
<td>Min Total Correlation obtained</td>
<td>1.27</td>
<td>2.40</td>
<td>1.91</td>
<td>1.60</td>
<td>2.57</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.47</td>
<td>0.85</td>
<td>1.06</td>
<td>1.87</td>
<td>2.10</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.19</td>
<td>0.58</td>
<td>0.89</td>
<td>1.88</td>
<td>2.54</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.13</td>
<td>0.14</td>
<td>0.07</td>
<td>1.43</td>
<td>1.88</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.34</td>
<td>0.52</td>
<td></td>
</tr>
</tbody>
</table>

5.3.7 Conclusions

An alternative method for siting a set of water level monitors in polders based on information quantities is presented. The first quantity is joint entropy, which evaluates the amount of information content that the set is able to collect; the second is total correlation, which evaluates the level of dependency or redundancy among monitors in the set. In order to find the most convenient locations to put the monitors from a large number of potential sites, a multiobjective optimization procedure (MOOP) was posed under different considerations: one that takes into account the costs of placing a completely new monitor, and another that considers the cost of placing monitors too close to hydraulic structures. In both cases, the joint entropy of the set is maximized and its total correlation is minimized. The costs are considered in terms of information theory units, for which additional terms $u*M$ and $q*v$ were introduced into the objective functions.

The following conclusions can be drawn:

- The information measures of Total Correlation and Joint Entropy are two conflicting quantities, because as the first improves (i.e., monitors are independent among them), the second deteriorates (i.e., monitors get less information content), and vice versa.
- The existing monitoring network, which reduces to the measurements taken at the pumping stations to control the switching of the pumps, is not optimal from the information theory point of view.

- The solutions for which total correlation was considered as a single objective (i.e., without joint entropy) are not satisfactory in terms of monitoring, because most of the monitors in the set are placed at sites with no information content. This means that points with no information content are the only ones that are able to add the least total correlation (independency) in a highly number of interdependent points. The Pareto solutions located at the extreme where the total correlation is minimum should, therefore,
be neglected. However, the maximization of joint entropy gives useful results, as the solutions found can cover between 82% and 85% of the total information content of the system by selecting fewer points than the originally proposed 9 points.

- The results obtained with the WMP method presented in section 5.2 are part of the optimal set of networks obtained by solving the MOOP posed in this section. The solution of the MOOP, however, has two main advantages over the WMP method: first, it gives a complete picture in terms of options to select a monitoring set; second, it allows adding constraints to the problem so a wider range of situations can be tested.

- In terms of the practical situations analyzed separately in this section, namely the financial constraint of having to place new monitors and the accuracy constraint of having to place them near hydraulic structures, it can be concluded that the best solutions in information units are numerically very similar (joint entropy for the first situation equal 4.18 bits and for the second situation equal to 4.04 bits), but spatially different.

### 5.4 Evaluation of the Monitoring Network of the Magdalena River

Until now only water level series have been used to estimate the Information Theory quantities for monitoring network design, due to the implications that they have for polder systems. For the case of natural streams, however, the analysis of discharge time series is more interesting than for water levels, since the effects of the tributaries on the behaviour of the river can be important. For this reason, two methodologies to design the discharge monitoring networks in rivers using concepts of information theory are presented in this section. The first methodology considers the optimization of Information Theory quantities and the second considers a new method that is based on ranking the different possible monitor combinations. The methodologies are tested for the Magdalena River in Colombia (see Chapter 4) in which the existing monitoring network is also assessed. In addition, the use of monitors at the tributaries is also explored. The ranking method is a promising way of finding the extremes of the Pareto fronts generated during the multiobjective optimization process.

#### 5.4.1 Description of the methodology

The ideal monitoring network would be composed of a set of gauges that provide the maximum information content and that are able to capture independent information. As previously mentioned, Total Correlation and Joint Entropy are two conflicting objectives: when the Joint Entropy of a set of variables increases, the Total Correlation decreases and vice versa. For this reason, the best location of $N$ monitors would be such that they simultaneously fulfil both objectives. A simpler mathematical formulation of the optimization problem than the one introduced for the case of Delfland in Eq. (5-5) is:

\[
\begin{align*}
\min \{C(X_1, X_2, \ldots, X_N)\} \\
\max \{H(X_1, X_2, \ldots, X_N)\}
\end{align*}
\]  

(5-8)

The optimization problem posed in Eq. (5-8) is solved using two different approaches: a) Multiobjective Optimisation with Genetic Algorithms (MOGA); 2) a ranked-based
greedy algorithm to optimize both objectives independently. Both approaches are described below.

5.4.1.1 Multi-objective optimization approach

One way of solving the problem posed in Eq. (5-8) is by looking at it as a multi-objective optimization problem (as in the section 5.3), which provides as a result a set of solutions that draws a Pareto front (efficient, non-dominated solutions). Such a front describes limits of what is possible in terms of decision criteria, and identifies how an improvement in one particular criterion is related to losses in other criteria. MOGA has been successfully used to solve water-related optimization problems (see e.g., Alfonso et al. 2010a; Barreto et al. 2009). A series of experiments were carried out to identify the optimal set of points for monitoring. In this section, NSGA-II, an elitist non-dominated sorting genetic algorithm for multi-objective optimization (Deb et al. 2002) is used.

5.4.1.2 Rank-based greedy algorithm

The second approach consist of a greedy algorithm that picks the best information-related gauge at a time is developed. The idea is to rank all the potential places to locate the monitors according to the variation in Joint Entropy and in Total Correlation, separately, caused by the selection of a new monitor. The algorithm requires the location of the first monitor to start, which could be defined as the monitor with the highest reduction in uncertainty, as suggested for instance by Krstanovic and Singh (1992). However, starting with the point that has the highest information content does not guarantee that the final set of monitors is the most informative, as is shown by the results. The second monitor is then chosen from the remaining set of monitors in such a way that it provides either the highest increment in Joint Entropy or the lowest increment in Total Correlation with respect to the first point. Flowcharts describing both situations are shown in Figure 5-21.

![Flowchart rank-based greedy algorithm for Joint Entropy](image1)

![Flowchart rank-based greedy algorithm for Total Correlation](image2)

Figure 5-21. Flowchart rank-based greedy algorithm for Joint Entropy (a) and Total Correlation (b)
5.4.2 Case study: Magdalena River, Colombia

The multiobjective optimization method and the ranked-based algorithm are applied to the monitoring network of the Magdalena River, the main river of Colombia, which runs for about 1,540 kilometres from South to North through the western half of the country to the Caribbean Sea. A detailed description of this case study is presented in Chapter 4.

Tributaries on the Magdalena River play an important role in many respects. The watershed and its main tributaries cover 257,400 km², which corresponds to 24% of the total surface of the national territory. The main tributaries, located mainly in the middle reach, have an important influence on the river’s behaviour in terms of discharge. Figure 5-22 presents the location of the most important cities near the river, the main tributaries and the available discharge and water level records for 1995.

Figure 5-22: Available hydrologic data records of discharges (Q) and water levels (h) at river stations for 1995.

The existing water-level gauges for the river were placed initially to support decision-making concerning local problems in the main populated areas, related to flood control and navigation, while keeping operation and maintenance costs low. However, from a global perspective, the information collected by these gauges is limited to supporting
decision and policy-making for navigation, flood control and other issues at other points of the river. Therefore, insight is needed on the design of a new monitoring network while evaluating the existing network in terms of its information content.

The selection of the parameter \( a \) in Eq. (5-1) is done in such a way that all of the tributaries included in the model have significant information content. That is, a big value of \( a \) would make small discharges insignificant in terms of information content because the rounding nature of the expression would transform the discharge series to a series with constant values. On the other hand, a too small value of \( a \) would make every discharge in the river have a similar information content, which is not convenient for our analysis. For these reasons, \( a \) was assumed to have a value of 200 m\(^3\)/s in order to include all the available discharges of the tributaries. A sensitivity analysis to show how results may change because of the selection of the parameter \( a \) is included at the end of this section.

5.4.3 Analysis of Results

The model output includes discharge time series for 181 calculation points along the main river (Loba branch) and 31 points for the Mompox branch. These time series are quantized using Eq. (5-1) adopting a value of 200 m\(^3\)/s for \( a \), which corresponds to the mean discharge of the smallest tributary with available data; so only the effects of inputs with this magnitude are captured by the entropy analysis.

The pre-design of the monitoring network for the Magdalena River was done using Information Theory to select a limited number of points from the 181 discharge points on the main channel where gauge devices are worth placing. As a first insight, the marginal entropy of each calculation point was estimated using Eq. (2-1) and a map of the entropy for the Magdalena River was prepared (Figure 5-23).

5.4.3.1 Analysis of the entropy map

Before presenting the solutions of Eq. (5-8) for the design of the monitoring network using the methods described above, the entropy maps (Figure 5-23, (a)) obtained for the discharge time series are compared to the map of discharges (Figure 5-23(b)). Firstly, entropy increases at points where the tributaries discharge into the river (see for example the rivers Miel, Negro, Nare, Sogamoso, Cimitarra, Cauca and the convergence of the branches Mompox and Loba in Figure 5-23(b)). The rivers Opón and Carare do not show any increment in the entropy, due to their relatively low influence in terms of discharge. The Mompox branch shows the same entropy along its channel, because there are no tributaries that flow into it. It is interesting to see that the lowest value of entropy occurs in this branch. This is because the discharge in this branch ranges from 400m\(^3\)/s to 1000m\(^3\)/s, so when applying Eq. (5-1) with \( a=200\text{m}^3/\text{s} \) the resulting quantized series have only four unique values in the frequency analysis and therefore only four sums are required to evaluate Eq. (5-8).
Secondly, entropy decreases when the wetlands interact with the river. As mentioned above, the wetlands act as a complex system of reservoirs that absorb the peak flows of the Magdalena River. For this reason, the discharge time series tends to be smooth, the range between minimum and maximum discharge is lowered and therefore entropy diminishes. Indeed, entropy is continuously increasing from upstream to downstream, until the wetland W1, just before the inflow of the Lebrija River. From this point to El Banco, entropy remains constant because no additional inflows exist. However, a big change in entropy takes place at El Banco, reducing it to the values reported after the inflows of the rivers Miel and Negro. This change is due to the connection of the Magdalena River to the wetland W2 (Figure 5-23), and the bifurcation of the main river into the Mompox and the Loba branches. It is clear that, after the bifurcation, the Loba branch contains, on average, a similar flow to that at San Pablo, or about 3800 m$^3$/s, highlighting again the effects of the wetlands. On the other hand, the effect of the wetland W3 on the entropy map is opposite that which is observed due to the first wetlands W1 and W2, since it adds entropy to the river. This effect is due to a third, minor bifurcation called Chicagua (not shown in Figure 5-23), which interacts with the so-called Momposina Depression, the “island” formed by the Mompox and Loba branches, in which water is discharged during high water levels events in the river. For the year 1995 the inflow was mainly from the depression to the Magdalena and therefore it was acting, on average, as an additional tributary. Although this additional discharge is not significant for the river (see Figure 5-23(a), between El Banco and Cauca inflow), it makes a difference in terms of entropy (see Figure 5-23(b)).
Thirdly, in the middle of the Loba reach, the biggest tributary, the Cauca River, flows into the Magdalena, greatly increasing its flow and also its entropy, which recovers some of the entropy absorbed by the wetlands W1 and W2. Additionally, although the wetland W4 acts in a similar way to W1 and W2 its effect is not apparent in the discharge map or the entropy map, because the influence of this zone is driven by the significant inflow from the Cauca River.

Finally, at the point where the branches Loba and Mompox converge, the entropy is a maximum. As there are no additional inflows or wetlands, this value remains constant until the most downstream point in Calamar.

5.4.3.2 Results using Multi-objective optimization approach

The multi-objective optimization problem posed in Eq. (5-8) is solved using the Non Sorted Genetic Algorithm, NSGA-II (Deb et al. 2002), for which the evolutionary parameters, namely, the number of populations and number of generations, must be specified. Additionally, the number of decision variables (number of monitors to be placed along the river) must be defined. In order to perform a sensitivity analysis of these parameters, a number of experiments were carried out, in which five different populations (P) and generations (G) were tested with the following combinations (P, G): (50, 20), (50, 50), (100, 20), (100, 50), (200, 50). Additionally, each experiment was carried out for a number of gauges from 6 to 9. The final solution was determined by selecting the best solutions (those with high Joint Entropy and low Total Correlation) from the five Pareto fronts. For comparison purposes, these solutions have been included in each Pareto front in Figure 5-24 as black dots.

![Figure 5-24: Solutions for multiobjective optimization approach. Black dots form the best Pareto front obtained by selecting the best points of the 5 combinations (P, G). Points A, B, C and D are selected for further analysis for 9 decision variables.](image-url)
Figure 5-24 it can be observed that the increment in the number of decision variables translates into a small increment in joint information and into a significant increase in redundant information. This means that new monitors will not add much more information content compared to what can be deduced from fewer monitors.

Figure 5-25 presents the locations of the solutions A, B, C and D shown in Figure 5-24 for the case of 9 decision variables on the entropy map (bits) of the Magdalena River. The redundancy of the solutions is evident, especially in the upstream part of the river. Additionally, Figure 5-26 presents the location of the solutions with the highest value of Joint Entropy for 6, 7, 8 and 9 monitors.

Several conclusions can be drawn from Figure 5-25. First, in general, the monitors are located where significant changes in entropy take place. Second, redundancy is reduced by adding monitors upstream, which do not add extra information content; on the contrary, joint information increases as more downstream monitors are added, with the consequent increment of their dependency. This confirms the trade-off between both information measurements. Finally, monitors are always selected at the Momposina Depression, especially where the wetlands have connections with the Magdalena and where the Cauca River discharges. The complex hydraulic conditions make the discharge change frequently along the river, leading to the increase of information content.
Moreover, Figure 5-26 shows the most informative solutions obtained for different numbers of monitors and how they are redistributed when an additional monitor is considered in the solution set. There is a regular distribution of the monitors for 6 and 7 decision variables, while for 8 and 9 monitors there is a tendency to locate the monitors upstream. This can be explained by recalling that all the tributaries, with the exception of the Cauca River, are located upstream (see Figure 5-22) so that the information collected at monitors located in this area provides insight into the state of the system downstream.

**5.4.3.3 Results using the ranked-based greedy algorithm**

The algorithms presented in Figure 5-21 were applied to the Magdalena River, in order to find the location of \( m \) number of monitors. In order to analyse the evolution of the monitoring network, experiments were carried out for \( m=5, 6, 7, 8 \) and 9. The algorithms were executed using as the starting point each of the 181 computational points of the model, generating 5 matrices (one for each \( m \)) with size 181 x \( m \).

- Ranking by joint entropy

The solution for the monitors with the maximum joint entropy was selected from the previously generated matrix; the locations of the monitors are plotted in Figure 5-27.
From Figure 5-27 it can be observed that the first monitor is located on the Loba branch, just before the convergence with the Mompox branch; this monitor, however, is not the one with the maximum information content (which is the second point from downstream to upstream). This means that starting with the monitor with the highest entropy does not guarantee that the final set of monitors has the maximum joint entropy.

The second monitor is located at the discharge of the Lebrija River and the connection to the wetland W2; the third monitor, that adds the maximum joint entropy to the previous set of two monitors, is placed after the discharge of the Nare river; the fourth monitor is located between the discharges of the rivers Carare and Opón and the fifth one is placed at the downstream part of the river, completing a set of five monitors with a Joint Entropy value of $H=8.4322$ bits. The solution for six monitors includes the same previous five locations in the same order of selection and adds the sixth at the place where the wetland W3 is connected, incrementing the Joint Entropy of the set to $H=8.4604$ bits. Similarly, the solution for seven monitors includes the previous six and adds the seventh at a place nearby the city of Berrío, downstream of the third monitor. This makes the Joint Entropy increase again to $H=8.4717$ bits.

Until now, every new selected monitor has been adding the maximum information content possible to the previous set and this monitor has been unique at every step. However, 20 different candidates arise for the monitor number eight and none of them provides any additional information content to the set of seven monitors, implying that...
further monitors are redundant. The location of the eight and ninth monitors (for which 148 different candidates arose) as shown in Figure 5-27 also confirms that these monitors are not worth selecting: they all congregate downstream repeating the information provided by the fifth monitor.

- Ranking by total correlation
The same exercise was performed for the algorithm presented in Figure 5-21 (b); results are shown in Figure 5-28. It can be observed that the first monitor is located at the connection to the wetland W3 and that the subsequent monitors were located at the upstream part of the river, looking for points with very low information content. This is because one way of reducing Total Correlation is by adding random variables with very low (or null) entropy (Alfonso et al. 2010c).

5.4.3.4 Comparison with the existing monitoring stations

The monitoring network formed by the existing stations on the Magdalena River, with flow data available for the year 1995, was evaluated from an Information Theory perspective. The set of 9 stations (Salgar, Berrio, San Pablo, Regidor, Peñoncito, El Banco, Magangué, Tacamocho and Calamar) has a value of Joint Entropy of \( H = 8.3808 \) bits and a value of Total Correlation of \( C = 34.7464 \) bits. The performance of this network can be compared to the results obtained for 9 variables using the multiobjective optimization approach and by the ranking approach in the Joint Entropy – Total Correlation space (Figure 5-30). It is observed that the set is not optimal (there exist other solutions that give better Joint Entropy and Total Correlation values).

Figure 5-28. Results obtained running the flowchart of Figure 5-21(b). Numbers represent the order in which each monitor was selected. The colour scale represents entropy (bits).
Comparison with monitors located at tributaries

From the practical point of view, discharge measurements are part of the navigation studies for the Magdalena River. In order to determine the water balance, these measurements are taken before and after the most important tributaries and at bifurcations such as the Mompox and Loba branches. In order to evaluate these locations from the Information Theory perspective, the value of the marginal entropy before and after the inflows of the eight tributaries included in the model is presented in Figure 5-29. It can be noted that the information content always increases after every inflow, with the exception of the Lebrija River, whose discharge is produced nearby the wetland W1. Therefore, the straightforward conclusion is to place monitors after the tributaries in order to get the maximum information content of the river.

However, a comparison between monitors located according to this analysis and the optimal solutions obtained with the multiobjective optimization method for 8 decision variables (bottom-right of Figure 5-24), reveals that the monitoring networks obtained considering the tributaries are sub-optimal; this suggests that the effect of the wetlands, typically ignored in the measurement campaigns due to the difficulties of monitoring in such a vast wetland area, the hundreds of small connections river-wetlands and the poor elevation data available, must be taken into account in order to understand the behaviour of the river better.

![Figure 5-29. Entropy values before, at and after the main tributaries](image)

To make a general comparison, the resulting monitoring set located taking into consideration the 8 tributaries of Figure 5-29 is included in the Total Correlation – Joint Entropy plane for 9 variables in Figure 5-30. It can be observed that both sets (before and after the tributaries) have a slightly better value of Joint Entropy than the existing monitors. Naturally, the Total Correlation cannot be evaluated in this graph because it is very sensitive to the number of monitors in place.
Figure 5-30. Solutions obtained by different methods, Total Correlation – Joint Entropy plane.

5.4.4 Sensitivity analysis of the parameter \(a\)

As previously mentioned, the selection of the parameter \(a\) in Eq. (5-1) may change the value of the entropy-related quantities. In order to analyse the implications of these changes, the entropy map presented in Figure 5-23 is redrawn for different values of \(a\) (see Figure 5-31).

It can be observed that entropy values decrease when the value of \(a\) increases, because the number of bins for the frequency analysis are fewer, and therefore the number of sums required to assess Eq. (2-1) is less. However, the relative value of the points with respect of the others in the same map is, in general, maintained regardless of the value of \(a\). Therefore the expressions (2-8) and (2-9) yield numerically different values, but basically the same locations are obtained. This can be seen in Figure 5-31, where the zones with high entropy are always between the discharges of the tributaries Cimitarra and Lebrija and also after the convergence of the branches Mompox and Loba. On the other hand, the zones with low entropy are located in the Mompox branch and at the upstream part of the river, before the discharge of the rivers Miel and Negro. It can also be observed that in the wetlands zone the entropy changes in a similar way between maps. This implies that the resultant monitoring networks generated with the presented methods do not change significantly when changing the value of \(a\). It must be noted, however, that an extreme, illogic value of \(a\) such as \(a=1\) or \(a=100.000\) m\(^3\)/s leads to useless constant entropy maps. It is recommended, therefore, that this value should be set between the lowest and the highest mean flow of the incoming tributaries of interest.
Figure 5-31. Entropy maps for different values of $a$, Eq. (5-1)

5.4.5 Conclusions and Recommendations

The entropy map for discharge in the Magdalena River shows that entropy increases at places where the tributaries flow into the river and diminishes at places where there exist connections to the wetlands.

The series of experiments carried out above gives rise to the following conclusions:

- The selection of high-entropy points for monitoring leads to redundant monitors and the selection of low-entropy points generates a final set with low information content. The conflicting nature of these Information Theory quantities promotes
the use of a multiobjective optimization approach. However, the selection of the final monitoring network selection is not straightforward if only the generated Pareto fronts are analysed, and it is still difficult to find an optimal solution that satisfies both criteria. In order to choose one point from the Pareto front, additional constraints are needed to determine the relative importance of joint entropy and total correlation. It is recommended that decision makers find these additional constraints by considering the requirements of water users.

- Seven monitors is the maximum number of monitors for which the Joint Entropy continuously increases along the Magdalena River, under the conditions in which the model was built. Additional monitors are fully redundant and do not add any further information content.

- The ranking-based methods are useful for finding the extremes of the Pareto fronts generated by the multiobjective optimization procedure and could be used in further research to normalise the information quantities and therefore to evaluate the solutions in a relative way. An interesting finding is that the initial monitor used to start the algorithms in Figure 5-21 plays a significant role in the Joint Entropy and the Total Correlation of the final set. Also, starting with the point with the highest entropy does not guarantee that the final set of monitors has the maximum information content.

- Although the existing monitoring stations were placed individually to fulfil the requirements of the cities without assessing the network as a whole, the performance of this set yields acceptable information content but there is a high redundancy between monitors. Moreover, its performance is similar to what is obtained if the monitors are located following the location of the tributaries, as is normally done during monitoring campaigns.

5.5 Conclusions

The distribution of the information content on a water system is driven by features that produce changes in the hydraulic conditions of the system. For the case of the polders of Pijnacker, the features that interfere with the distribution of the information content are the weirs and the pumps; for the case of the Magdalena River, it is the incidence of its wetlands and tributaries. However, this does not mean that a monitoring network configured by placing monitors at the hydraulic structures in the first case or at the tributaries in the second case will be optimal from an Information Theory perspective.

In all the experiments, the trade-off between Total Correlation and Joint Entropy is evident. However, a large Joint Entropy can be preferred over a small Total Correlation. This is because low-dependent set of monitors are also the least informative. In this case, therefore, it does not make sense to have a completely independent set of monitors if they do not provide enough information content individually.

The location with the highest information content of the system is usually the most dependent of the remaining locations. This implies that once this point is selected, it is very difficult to find a second point that is informative and at the same time is independent. This also explains why only a few monitors are needed in spite of having
many possibilities (8 out of 1520 potential monitors in place for the Pijnacker region and 7 out of 181 for the case of the Magdalena River).

The sensitivity of any discretization-based criteria for the assessment of probability distributions is a well-known difficulty that greatly affects the estimation of entropy-related quantities. However, these effects are negligible for the methods determining monitor location due to the relative nature of the developed methods.
Chapter 6
Value of Information for monitor location

This chapter presents a novel approach for locating monitors in a water system using the concept of Value of Information (VOI). This concept takes into account three main factors: 1) the belief that the decision-maker has about the state of the water system before having any information; 2) the consequences associated with the decision of having to choose among several possible actions given the state of the water system; and 3) the evaluation and update of new information when it becomes available. The methodology uses water level time series generated by hydrodynamic models at every computational point, each one being a potential monitor site. The method is tested in two case studies of completely different nature: the Magdalena River in Colombia, and a polder system in The Netherlands. It is shown that the methodology can be used as a complementary approach to existing methods that focus the monitor location problem exclusively on information-theory.

This Chapter is organised as follows: first, the main considerations are presented in the introduction, followed by a section that defines the variables for the VOI estimation. Then an explanation of how monitor locations can be valued according to the VOI theory is presented. The approach to locating monitors is then explained for the case of one, two and three monitors and the generalization of the method for n monitors. Subsequently, the methods are applied to the Magdalena River and to the Pijnacker polders, after a brief test in a hypothetical, simple case. Finally, the conclusions are presented for this Chapter.

6.1 Introduction

Two main considerations are taken into account when developing the methods presented in this Chapter: the use of a model as a data generator, in a similar way as presented in Chapter 5, and the use of the generated data to estimate the probabilities required for the VOI calculation.

Firstly, in this Chapter the model is used to generate time series from which the probabilities required to assess the Value of Information are estimated. The model is required because available measurements are generally limited to a few points making them insufficient to draw conclusions from their analysis. In contrast, the model
generates a dense set of points. As presented in Chapter 5, every calculation point within a model is considered as a potential location for a monitoring point within a water system. The details of the model used for the case of the polders of Pijnacker and for the case of the Magdalena River in Colombia have been presented in Chapter 3 and in Chapter 4 respectively.

Secondly, this chapter introduces a procedure to estimate the prior beliefs, \( \pi_s \), and the conditional probabilities, \( q_{m,s} \), used in the VOI estimation. The assessment of these parameters is difficult, because the probabilities before and after receiving the new information are not known. The data generated by the model is used to estimate such probabilities with the procedure explained in the following section.

### 6.2 Definition of variables for VOI estimation

As reviewed in Chapter 2, VOI is a function of the consequences \( c_{as} \), of taking an action, \( a \), given a particular state, \( s \), of the prior probability \( \pi_s \), or the belief before the acquisition of additional information, and of the conditional probability \( q_{m,s} \), of receiving the message, \( m \), given the state, \( s \). This means that three sets of data are needed to estimate the Value of Information: the set of actions \( a \), the set of possible states \( s \) and the set of messages \( m \) that the monitors provide.

First, the set \( a = (a_1, a_2, ..., a_A) \) contains \( A \) actions that are available for the decision maker in order for him \(^2\) to deal with the state of the system, for example, by turning a pump on, by releasing a warning or simply doing nothing. Second, the set \( s = (s_1, s_2, ..., s_S) \) contains \( S \) possible states of the system, namely, flooding or normal water levels, that should be defined for each point of the system. Third, the set \( m = (m_1, m_2, ..., m_M) \) contains \( M \) messages that the information service will provide to the decision maker as an indication of the possible state of the system. Examples of such messages can be “Danger” or “Relax”. Two actions, two states and two messages are considered in this Chapter for simplicity.

#### 6.2.1 Estimation of the probabilistic variables \( \pi_s \) and \( q_{m,s} \)

In the first place, the prior probabilities associated with the possible states are estimated using the vector \( \pi_s \), shown in Table 6-1 for the case of two possible states \( s_1 \) and \( s_2 \) that can occur in the given water system. Two values are defined: 1) Number of states \( s_i \) at point \( i \), referring to the number of times the state \( s_i \) has been repeated in the history of the point \( i \) in the water system; 2) Number of records, being the length of the generated time series at point \( i \), which is generally a constant for all points in the system. Consequently, each element in the \( \pi_s \) vector contains the relative frequency of each state, which can be regarded as the probability of a particular state occurring at each point of the system, and in this respect may be treated as the knowledge the decision maker has about his system for a given state.

\(^2\) We imply both male and female when using him/his/he to refer to the decision-maker.
Table 6-1. Definition of the vector $\mathbf{r}_s$ for two possible states of a water system

<table>
<thead>
<tr>
<th>State</th>
<th>$\mathbf{r}_s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$: State 1</td>
<td>Number of states $s_1$ at point $i$ / number of records</td>
</tr>
<tr>
<td>$s_2$: State 2</td>
<td>Number of states $s_2$ at point $i$ / number of records</td>
</tr>
</tbody>
</table>

In the second place, the conditional probabilities $q_{m,s}$ of receiving the message $m$, given the state $s$, are estimated by checking the relative frequency of the situations presented in Table 6-2, where $x$ is the location at which messages are produced taking into account the state at the location, and $y$ is any other point in the system.

Table 6-2. Possible situations of messages at $x$ for given states at $y$ for the case of two possible states

<table>
<thead>
<tr>
<th>Situation</th>
<th>Message at $x$</th>
<th>State at $y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$m_1$</td>
<td>State 1</td>
</tr>
<tr>
<td>2</td>
<td>$m_1$</td>
<td>State 2</td>
</tr>
<tr>
<td>3</td>
<td>$m_2$</td>
<td>State 1</td>
</tr>
<tr>
<td>4</td>
<td>$m_2$</td>
<td>State 2</td>
</tr>
</tbody>
</table>

From Table 6-2, bearing in mind that the messages $m_1$ and $m_2$ describe the states $s_1$ and $s_2$ respectively, it is clear that:
- the messages are correct when situations 1 and 4 happen;
- there is a type I-error in situation 3;
- there is a type II error in situation 2.

For example, consider that $m_1 =$ “Danger” is used to describe the state $s_1 =$ ”Flood” and $m_2 =$ ”No Panic” is used to describe the state $s_2 =$ ”Normal”. In the situation 3 in Table 6-2, the message $m_2$ incorrectly announces that there is no problem, while in the situation 2 the message $m_2$ incorrectly announces that a flood event is occurring.

Therefore, the conditional probabilities $q_{m,s}$ may be estimated as stated in Table 6-3, where the value $Num$ situation $i$ is the number of times a situation $i$ (Table 6-2) happened in the history of the time series and $Num$ states $s$ is the number of times the particular state $s$ occurred.

Table 6-3 Definition of conditional probabilities $q_{m,s}$ according to the situations presented in Table 6-2.

<table>
<thead>
<tr>
<th>$q_{m,s}$</th>
<th>$m_1$: Message 1 (at $x$)</th>
<th>$m_2$: Message 2 (at $x$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$: State 1 at $y$</td>
<td>$Num$ situation 1 / $Num$ states $s_1$</td>
<td>$Num$ situation 3 / $Num$ states $s_1$</td>
</tr>
<tr>
<td>$s_2$: State 2 at $y$</td>
<td>$Num$ situation 2 / $Num$ states $s_2$</td>
<td>$Num$ situation 4 / $Num$ states $s_2$</td>
</tr>
</tbody>
</table>

It can be observed that the elements in each row sum up to one. However, a row will have a value of zero if the corresponding state does not occur. This means that a monitor at $x$ makes sense only if it is able to say something about the state at $y$. 

105
6.2.2 Definition of the consequences $c_{as}$

The consequences, $c_{as}$, of taking an action, $a$, given a particular state $s$, form a matrix that contains the costs associated with having chosen to perform an action according to the state the decision maker thinks is happening. Table 6-4 presents an example of this matrix for the case of two actions and two states.

Table 6-4. Definition of the $C_{as}$ matrix.

<table>
<thead>
<tr>
<th>$C_{as}$</th>
<th>$a_1$: action 1</th>
<th>$a_2$: action 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$: State 1</td>
<td>Cost of doing $a_1$ when $s_1$</td>
<td>Cost of doing $a_2$ when $s_1$</td>
</tr>
<tr>
<td>$s_2$: State 2</td>
<td>Cost of doing $a_1$ when $s_2$</td>
<td>Cost of doing $a_2$ when $s_2$</td>
</tr>
</tbody>
</table>

Naturally, this matrix depends on the type of water system; its definition is not straightforward because of the hypothetical character of the damages under different scenarios, especially for extreme states. However, this consequence matrix can be built according to the judgment of water board experts (see, e.g., van Andel 2009 p. 116-118).

In order to clarify the calculations, a numerical example of the procedure to calculate the Value of Information is shown next.

**Procedure**

Following the flowchart presented in Figure 2.3, a numerical example of the calculation procedure is presented assuming that the calculation of Table 6-1 and Table 6-2 for a given point in the system yields, respectively:

$$\pi_s = \begin{bmatrix} 0.925 \\ 0.075 \end{bmatrix}, \quad q_{m,s} = \begin{bmatrix} 0.75 & 0.25 \\ 0.10 & 0.90 \end{bmatrix}$$

Suppose also that the matrix of consequences $C_{as}$ (explained in detail further) is:

$$c_{as} = \begin{bmatrix} -5 & -100 \\ -30 & 0 \end{bmatrix}$$

First, the action that would have been chosen without information, $u(a_0, \pi_s)$, is calculated as the maximum utility given by performing each action $a$ based on the prior beliefs $\pi_s$:

$$u(a_0, \pi_s) = \max \left\{ \sum_{s} \begin{bmatrix} -5 & -100 \\ -30 & 0 \end{bmatrix} \begin{bmatrix} 0.925 \\ 0.075 \end{bmatrix} \right\} = -6.875$$

Then, the posterior probabilities are calculated and a total of $M*S$ values are obtained (one for each combination of states $s$ and messages $m$):

$$\pi_{s,m} = \frac{q_{m,s} \pi_s}{\sum_{s} q_{m,s} \pi_s} = \begin{bmatrix} 0.925*0.75 & 0.925*0.25 \\ 0.075*0.10 & 0.075*0.10 \end{bmatrix} = \begin{bmatrix} 0.989 & 0.774 \\ 0.011 & 0.226 \end{bmatrix}$$

Next, the expected utility, the probability-weighted average of the utilities of the associated consequences, is estimated:
Chapter 6 - Value of Information for monitor location

\[
\begin{align*}
    u(a, \pi_{s,m}) &= \sum_s c_{as} \pi_{s,m} = \\
    &= \begin{bmatrix}
      -5 & -100 \\
      -30 & 0
    \end{bmatrix} \begin{bmatrix}
      0.989 \\
      0.011
    \end{bmatrix} = \\
    &= \begin{bmatrix}
      -10.558 \\
      -77.406
    \end{bmatrix}
\end{align*}
\]

The decision-maker will choose the action that gives the maximum utility for all possible states, so:

\[
    u(a_m, \pi_{s,m}) = \max_a \left\{ \begin{bmatrix}
      -5 & -100 \\
      -30 & 0
    \end{bmatrix} \begin{bmatrix}
      0.989 \\
      0.011
    \end{bmatrix} \right\} = \begin{bmatrix}
      -5.267 \\
      -98.931
    \end{bmatrix}
\]

Then, the value of each message is calculated as the difference of the utilities of the action \(a_m\) that is chosen given the message \(m\), and the utility of the action \(a_0\) that would have been chosen before additional information:

\[
    \Delta_m = u(a_m, \pi_{s,m}) - u(a_0, \pi_{s,m}) = \begin{bmatrix}
      -5.267 \\
      -98.931
    \end{bmatrix} - \begin{bmatrix}
      -6.875 \\
      -1.608
    \end{bmatrix} = \begin{bmatrix}
      -1.608 \\
      -97.323
    \end{bmatrix}
\]

Finally, the Value of Information is the expected utility of the new information:

\[
    VOI = \sum_m q_m \Delta_m = (0.925*0.75 + 0.075*0.10)*1.608 + (0.925*0.25 + 0.075*0.90)*(-3.774)
\]

\[
    VOI = 0
\]

In this case, the VOI with the given \(q_{m,s}\) is zero because the action that would have been chosen with the additional information does not have an expected utility. This is due to several factors. First, the decision-makers’ prior belief has a tight distribution (his confidence about the state of the system is very high). Second, the perceived quality of the information service given by the matrix \(q_{m,s}\) makes it useless. Certainly, 25% of the time the message will incorrectly reject the true state (there is a flood and the message does not show it), while 10% of the time it will incorrectly reject the false state (there is no flood but the message says there is). Third, the outcome of the message is in line with what the decision maker believes.

![Figure 6-1. Variation of the Value of Information when changing the prior probability \(\pi_x\) and the conditional probabilities \(q_{m,s}\) for the consequence matrix shown in Table 6-6.](image)
If the same exercise is repeated assuming that the decision maker is completely ignorant of the state of his/her system (this is $\pi_0 = 0.5$), then VOI = 1.62. Additionally, if the quality of the information service is improved (i.e., $q_{m,s} = [1 \ 0 \ 0 \ 1]$), then VOI yields 15.00. The variations of the VOI for the consequence matrix shown in Table 6-6 are presented in Figure 6-1, where the conditional probabilities $q_{m1,s1}$ and $q_{m1,s2}$ are presented on the x-axis and the prior probabilities $\pi_0$ on the y-axis. It can be observed that the maximum VOI is 22.00 for the given consequence matrix, which is achieved only if the quality of the message is perfect.

6.3 Value of the location for one monitor

For the subsequent discussion, a series of propositions are presented, taking into account that $x$ is the location in a water system where a monitoring device is placed to provide messages about the possible states of any point, $y$, on which a decision maker bases his water management related decisions. For simplicity, it is considered, in principle, that the water system is a linear canal.

- **Proposition 1**

  The value a decision-maker is willing to pay for a monitor located at $x$ to know the state at any other point $y$ is given by

  \[
  V_x(y) = VOI(c_{st}, \pi_0, q_{m,s,y})
  \]

  $V_x(y)$ is a single value if $y$ is one point and is a vector if it is calculated for all points $y$ in the water system (see Figure 6-2).

- **Proposition 2**

  In order to know the state of the world at point $x$, the most convenient point to place a monitor is precisely (and obviously) the point $x$.

  The proposition 2 implies that $V_x(x)$ is a maximum, that is, there exist no other locations where a monitor can provide a higher value about the state of the system at $x$ (see Figure 6-2). This is because the conditional probabilities $q_{m,s,x}$ estimated as in Table 6-3 yield the unit matrix.

  Moreover, if $V_x(y)$ is calculated individually for all points $y$ in the system, then the curve for $V_x$, shown in Figure 6-3(a) is obtained. The curve has a maximum at $x$ and decreases progressively as $y$ moves away from $x$, because some of the messages produced at $x$ do not coincide with the states at $y$ and the conditional probabilities $q_{m,s,x}$ do not give unity for the corresponding matrix. Therefore, the mean Value of Information that the location $x$ gives about the state of the entire system, $VOI_x$, is defined as the area below the $V_x$ curve along the canal with length $L$:

  \[
  VOI_x = \int V_x(y) \, dl \quad \text{(6-1)}
  \]
However, from the practical viewpoint, the water system will always have a finite set of \( N \) points and then the resulting curve \( V_x \) is a discrete one (Figure 6-3(b)); therefore, the mean Value of Information that the location \( x \) gives about the state of the entire system can be defined as:

\[
VOL_x = \sum_{y=1}^{N} V_x(y) \tag{6-2}
\]
6.4 Value of the locations for two monitors

The best place to locate a monitor to know the state at point \( y \) is the point \( y \) itself, and any other monitor located at \( a \) and used to know the state of the system at \( y \) has a value \( V_a(y) \), which is lower than \( V_y(y) \). Now, if an individual needs an additional monitor located at \( b \) to know better the state at the same point \( y \) then, first of all, he should not pay more than \( V_b(y) \) for it. Additionally, as the existing monitor at \( a \) is already giving some information about \( y \) with value \( V_a(y) \), then the maximum value the individual should be willing to pay for the additional monitor located at \( b \) is \( V_b(y) - V_a(y) \); see Figure 6-4. This approach is in agreement with the fact that the value an individual is willing to pay for placing additional monitors to know the state at the same point \( y \) is progressively lower as new monitors are added.

Naturally, the difference \( V_b(y) - V_a(y) \) should be positive, and should not be bigger than \( V_y(y) \); in other words, the ideal location \( b \) to place a second monitor is such that \( V_b(y) - V_a(y) = V_y(y) \). This is the basis of the approach described below.

6.5 Selection of monitor locations based on VOI

In this section the value of a monitor as described in the previous section is used to develop the method for selecting the best locations for monitoring a water system. The method is presented in steps involving one monitor at a time.

6.5.1 Locating one new monitor

Proposition 1 introduced in Section 6.3 states that the maximum value a decision-maker is willing to pay for a monitor to be located at \( x \) in order to know the state at point \( y \) is given by \( V_x(y) \). If the exercise is repeated allowing the point \( x \) to move along the water system, a family of curves \( V_x \) (and therefore a vector of \( \text{VOI}_x \)) is obtained. It follows that the best place to locate one monitor is where it can provide messages about the state of the maximum number of points in the system; therefore such a monitor should be located at \( x \) where it has the largest \( \text{VOI}_x \), namely:
Chapter 6 - Value of Information for monitor location

\[
\max \{VOI_x\} = \max \left\{ \int V_x(y) \, dl \right\}
\]  

(6-3)

Intuitively, the monitor located at \(x_3\) in Figure 6-5 is the one that is more valuable to capture the state of the largest part of the system, even though \(V_{x_1}\) has the maximum VOI among the three monitors considered. For this reason, the best monitor is one that provides the maximum area (or maximum averaged area) below its curve \(V_x\).

Figure 6-5. Selection of the best monitors out of three possibilities

6.5.2 Locating two new monitors

Following the reasoning presented in Section 6.4, the simultaneous location of two monitors \(a\) and \(b\), is an optimization problem that must be solved as presented in Eq.(6-4). This procedure looks for two points in the system such that the VOI of the monitor \(a\) plus the positive area between the curves \(V_a\) and \(V_b\) \((A_{ab})\) is a maximum. (see Figure 6-6).

\[
\max \{VOI_a + A_{ab}\} \\
A_{ab} = \int (V_b - V_a) \, dl; \{V_b(y) - V_a(y)\} > 0
\]  

(6-4)

Figure 6-6. VOI-related areas to optimise the monitor locations \(a\) and \(b\), Eq. (6-4)
6.5.3 Locating three new monitors

The same procedure for two monitors is used in simultaneously locating three monitors $a$, $b$ and $c$. The mathematical expression for the optimization problem is presented in Eq. (6-5), and depicted in Figure 6-7.

$$\text{max} \left\{ VOI_a + A_{ab} + A_{abc} \right\}$$
$$VOI_a = \int V_a \, dl$$
$$A_{ab} = \int (V_b - V_a) \, dl; \{V_b(y) - V_a(y)\} > 0$$
$$A_{abc} = \int (V_c - V_b - V_a) \, dl; \{V_c(y) - V_b(y) - V_a(y)\} > 0$$ 

$$\text{Figure 6-7. VOI-related areas to optimise the monitor locations a, b and c, Eq. (6-5).}$$

6.5.4 Locating $N$ new monitors

The generalization of the optimization problem for the case of $N$ monitors is given by Eq. (6-6):

$$\text{max} \left\{ VOI_a + A_{ab} + A_{abc} + \ldots + A_{abc...N} \right\}$$
$$VOI_a = \int V_a \, dl$$
$$A_{ab} = \int (V_b - V_a) \, dl; \{V_b(y) - V_a(y)\} > 0$$
$$A_{abc} = \int (V_c - V_b - V_a) \, dl; \{V_c(y) - V_b(y) - V_a(y)\} > 0$$
$$A_{abc...z} = \int (V_N - \ldots - V_c - V_b - V_a) \, dl; \{V_N(y) - \ldots - V_c(y) - V_b(y) - V_a(y)\} > 0$$

The application of the procedure for monitor location is described in the case studies given in the following section.
6.6 Case studies

In order to test VOI approach for siting monitors, experiments in three different water systems are presented here. The first experiment is carried out for a simple, hypothetical canal that is controlled by a pump at one end of its reach. Subsequently, the method is applied to two real but very different water systems in nature: the Magdalena River, the most important river of Colombia, and the canal network of the Pijnacker polders in The Netherlands.

6.6.1 Canal and pump

Consider a canal that receives the drainage of a big polder area. At the end of the canal there is a pump station to drain the water out of the polder. The pump operation is based on water levels measured at its suction side. In order to see how the VOI approach works, a particular flood level is defined, which changes linearly from the upstream end, where flooding always occurs, to the downstream end where flooding never occurs (see Figure 6-8). The actions, states and messages to evaluate the value of information are presented in Table 6-5.

![Figure 6-8. Definition of flood levels for the canal-pump experiment compared to the minimum, mean and maximum water levels obtained by the model](image)

<table>
<thead>
<tr>
<th>Calculation point</th>
<th>Elevation (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>0.4</td>
<td>0.4</td>
</tr>
<tr>
<td>0.6</td>
<td>0.6</td>
</tr>
<tr>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1.2</td>
<td>1.2</td>
</tr>
<tr>
<td>1.4</td>
<td>1.4</td>
</tr>
<tr>
<td>1.6</td>
<td>1.6</td>
</tr>
<tr>
<td>1.8</td>
<td>1.8</td>
</tr>
</tbody>
</table>

Table 6-5. Definition of actions, states and messages for the canal-pump case

<table>
<thead>
<tr>
<th>Actions to choose from</th>
<th>Possible ‘States of the world’</th>
<th>Messages (from the monitor at x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1$: Pump On</td>
<td>$s_1$: Flood (anywhere)</td>
<td>$m_1$: Danger</td>
</tr>
<tr>
<td>$a_2$: Do Nothing</td>
<td>$s_2$: No flood (anywhere)</td>
<td>$m_2$: Normal</td>
</tr>
</tbody>
</table>

The consequences, $c_{ars}$, of taking an action, $a$, given a particular state $s$, are assumed to be constant for every point of the system. Due to the lack of data, it is also assumed that, on the one hand, the cost of releasing a warning when there is flooding is -$5 (the costs
associated with communication and mobilization) and is -$30 if there is no flooding (the damage costs of having turned the pump on when no flooding occurs). On the other hand, the cost of doing nothing when there is flooding is -$100 (a relatively high cost due to the disaster associated with a flood that is made worse by the fact that the pump was not turned on) and is $0 if there is no flooding (see Table 6-6). This is the same consequence matrix used in the numerical example shown at the end of section 6.2.

<table>
<thead>
<tr>
<th>( s \times a )</th>
<th>( a_1 )</th>
<th>( a_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( s_1 )</td>
<td>-5</td>
<td>-100</td>
</tr>
<tr>
<td>( s_2 )</td>
<td>-30</td>
<td>0</td>
</tr>
</tbody>
</table>

### Results

Before proceeding with the location of the monitors, an analysis of the value that a monitor located at \( x \) provides when \( x \) changes along the canal, is presented. For this analysis, the prior probabilities, or the beliefs about the states before receiving additional information (estimated as shown in Table 6-1), are depicted in Figure 6-9:

![Figure 6-9. Prior beliefs \( \pi_i \) estimated with the Table 6-1](image)

Additionally, the \( VOI_x \) presented in Eq.(6-2), is calculated for each point in the system. The results are shown in Figure 6-10, where the flood level definition and the maximum and minimum water levels are also presented:

From Figure 6-10 it can be seen that the calculation points which are always flooded (1 to 4) and the calculation points which are never flooded (93 to 102) have no value, because the decision-maker’s experience (replicated by the analysis of each time series as shown
in Table 6-1) shows that he is completely certain about the state of the system in these areas ($\pi = [0;1]$ upstream and $\pi = [1;0]$ downstream); see Figure 6-9.

Additionally, the point with the maximum $VOI_x$ is located between points 40 and 43, where $\pi$ is very close to the vector $[0.5;0.5]$ (Figure 6-9), implying that the decision maker is completely uncertain about the state of the system at these points.

**Location of one monitor**

The solution of Eq. (6-3) yields point 43 as the location that provides the best description of the state of the system. Its curve $V_{43}$ is shown in Figure 6-11.
It can be observed that this point is not able to produce reliable messages about the state of the system upstream of point 31 and downstream of point 81. The curve has a constant value at the top because at those points the expected value of the action taken without information $a_0$ increases and the expected value of the action $a$ to be taken after receiving the new information decreases, so their difference yields the same numerical value.

**Location of two monitors**

Similarly, the solution of the Eq. (6-4) yields points 37 and 57 as the pair of points that together are the most valuable to describe the state of the water system. The curves $V_{37}$ and $V_{57}$ are presented in Figure 6-12. The point 43 selected for the case of one monitor is not selected, because, 37 and 57 together provide reliable messages about the state of the entire system. It can be seen that $V_{57}(57) = 19.30$ is larger than $V_{43}(43) = 15.36$, but the former provides information for more points in the system than the latter. This fact, however, is modified by the addition of the second monitor 37, with a maximum value of $V_{37}(37)=13.51$.

![Figure 6-12. V curves for the monitors 37 and 57, after solving Eq. (6-4)](image)

**Location of three monitors**

The solution of Eq. (6-5) yields points 4, 38 and 65 and Figure 6-13 shows their $V$ curves. The selection of point 65 makes sense because $V_{65}(65)= 21.69$, a value that is very close to the maximum possible for the considered consequence matrix (see Figure 6-1). Point 38 does the same job as the previously selected point 37, which is to provide information about the state of the system where there is high uncertainty (in the middle of the reach). Finally, the point 4 is selected as being complementary to points 65 and 38. This point covers the area of the canal where there is high certainty, and this explains why its maximum value $V_4(4)$ is so low (0.14).
In this case study the location of flow monitors is considered. Details of the Magdalena River and the developed hydrodynamic model are presented in Chapter 4. The actions, states and messages required for the VOI analysis are developed here. First, the set of actions are:

- \( a_1 \): Release a flood warning;
- \( a_2 \): Do nothing.

Second, the set of possible states are:

- \( s_1 \): Flooding;
- \( s_2 \): Normal.

Finally, the messages the monitors can provide are:

- \( m_1 \) = “Danger”;
- \( m_2 \) = “Relax”.

For comparison purposes, the consequences, \( c_{as} \), shown in Table 6-7 are the same as those used in the canal-pump example.

**Table 6-7. Consequences of doing action \( a \) given state \( s \) (costs units)**

<table>
<thead>
<tr>
<th>( c_{as} )</th>
<th>( a_1 )</th>
<th>( a_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( s_1 )</td>
<td>-5</td>
<td>-100</td>
</tr>
<tr>
<td>( s_2 )</td>
<td>-30</td>
<td>0</td>
</tr>
</tbody>
</table>

So far as costs are concerned, the cost of releasing a warning when there is flooding is -$5 (the costs associated with communication and mobilization) but is -$30 if there is no flooding (the political costs of having mobilized people under a false alarm). Correspondingly, the cost of doing nothing when there is flooding is -$100 (a high cost due to the disaster associated to the flood that was not warned), but is $0 if there is no flooding.

**Celerity and lagged time series**

In the case of the canal and pump presented in Section 6.6.1 the definition of the conditional probabilities \( q_{ms} \) (Table 6-3), is based on the premise that a message at point \( x \) may have some value for providing the state of the system at a point \( y \), both states occurring simultaneously. Although this is a correct principle for flat water systems, an important additional consideration for the estimation of the conditional probabilities \( q_{ms} \),
for the Magdalena River, must take into account the motion of the kinematic wave in the river.

Indeed, in a non-flat system such as a river, a message produced at time \( t = t_0 \) from a monitor located at \( x \) might provide more value for indicating the state of the system at a downstream point \( y \) some time in the future (\( t = t_0 + \Delta t \)), than the value it might provide at time \( t = t_0 \). Therefore, in order to account for this issue, we define the matrix \( q_{m,s} \) as follows:

Table 6-8. Definition of the situations for estimation of \( q_{m,s} \) for the Magdalena River case

<table>
<thead>
<tr>
<th>Situation</th>
<th>Message at ( x ) at time ( t_0 )</th>
<th>State at ( y ) at time ( t_0 + \Delta t )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>“Danger”</td>
<td>“Flooding”</td>
</tr>
<tr>
<td>2</td>
<td>“Danger”</td>
<td>“No Flooding”</td>
</tr>
<tr>
<td>3</td>
<td>“Normal”</td>
<td>“Flooding”</td>
</tr>
<tr>
<td>4</td>
<td>“Normal”</td>
<td>“No Flooding”</td>
</tr>
</tbody>
</table>

The new matrix \( q_{m,s} \) is shown in Table 6-9.

Table 6-9. Definition of conditional probabilities \( q_{m,s} \) according to the situations presented in Table 6-8.

<table>
<thead>
<tr>
<th>( q_{m,s} )</th>
<th>( m_1 ): Danger (at ( x )) at time ( t_0 )</th>
<th>( m_2 ): Normal (at ( x )) at time ( t_0 + \Delta t )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( s_1 ) Flood at ( y ) (at time ( t_0 + \Delta t ))</td>
<td>Number of occurrences of situation 1 / Number of “Flooding” at ( y )</td>
<td>Number of occurrences of situation 3 / Number of “Flooding” at ( y )</td>
</tr>
<tr>
<td>( s_2 ) No Flood at ( y ) (at time ( t_0 + \Delta t ))</td>
<td>Number of occurrences of situation 2 / Number of “No Flooding” at ( y )</td>
<td>Number of occurrences of situation 4 / Number of “No Flooding” at ( y )</td>
</tr>
</tbody>
</table>

Three of the situations described in Table 6-9 are depicted in Figure 6-14, where the position of a kinematic (flood) wave is shown at different times. In brief, the quality of the message provided by \( x \) at \( t = t_0 \) is successful in describing the state of the system at \( y \) at a time \( t = t_2 \); however, it is not successful for describing the state of the same point at a time \( t = t_1 \).

It can be seen that the definition of the critical level for flooding (i.e., the threshold that defines the possible states of the system), has an important effect on the quality of the messages that are produced by \( x \) about the state of the system at \( y \).

From a logical point of view, Table 6-9 is estimated by lagging the time series obtained from the hydrodynamic model at all computational points. The lag time used is equal to the travel time \( T \) of the kinematic wave, which is estimated by:
\[ T = \frac{dx}{c} \]  

(6-7)

where \( dx \) is the distance between two consecutive computational points and \( c \) is the celerity of the kinematic wave.

**Definition of thresholds for the states of the Magdalena River**

The states of the Magdalena River are defined for different thresholds in terms of percentiles at each computational point. The purpose of this definition is to let the decision maker choose what type of state he wishes to monitor. For instance, a threshold of 80% means that the set of monitors to be placed is needed to provide the best information about the flows above the 80% threshold of the time series for the flow at each point.

In this way, a sensitivity analysis of the value of information according to different thresholds can be investigated. However, in practice, real thresholds, defined by the water boards and decision makers should be used.

**Results**

The results are shown for the location of one, two and three monitors in the Magdalena River, under different celerity values and different state definition thresholds. Before
presenting the results, it is worth analysing the mean Value of Information $VOL_t$ estimated with Eq. (6-2), because it provides a picture of the most important zones in terms of monitoring the state of the entire system. Figure 6-15, Figure 6-16 and Figure 6-17 have been prepared using state thresholds of 80%, 50% and 20% respectively, and for different values of celerity.

First, from Figure 6-15 it can be observed that the tributaries and the wetlands affect the variation of $VOL_t$ in different ways; in some cases they increase VOI (for example the rivers Negro, Nare and Lebrija and wetland $W_4$) and in other cases they decrease it (for example the rivers Sogamoso and Cauca and the wetlands $W_1$ and $W_2$). In general, the region between the Lebrija River and the wetland $W_2$ is the one that provides the highest quality of the messages about the state of the entire system. This is consistent with the fact that this zone divides the river into two reaches: an upstream reach, where the majority of the tributaries flow into the river, and the wetland reach, where the river slope decreases sharply to a low value.

![Figure 6-15](image.png)

Figure 6-15. Mean Value of Information estimated with Eq. (6-2) for different values of celerity in the Magdalena River, using a state threshold definition of 80%.

Second, from Figure 6-16 for a state threshold definition of 50% it can be observed that the mean value of information is more sensitive to smaller disturbances in discharge; this is especially evident in the upstream reach, in particular for the rivers Cocorná (between the rivers Miel and Nare), Regla (between the rivers Nare and Carare) and Opón (between the rivers Carare and Sogamoso). Similarly, the Cesar River becomes important in the wetland zone, located after the wetland $W_2$ (see Figure 6-16 and Figure 4-2).
Finally, for a state threshold of 20% (see Figure 6-17), the \( VOI_x \) curves for different celerity values show that the downstream reach of the river (after the connection of the Mompox branch) becomes as important as the zone between the Lebrija River and the wetland \( W_2 \). Note that the mean value of information at the wetland zone between \( W_2 \) and \( W_3 \) decreases significantly, implying that, on average, almost any location is unable to provide reliable messages about the state of the system in this particular zone. Conversely, any monitor located within \( W_2 \) and \( W_3 \) will be unable to provide reliable messages about the state of the entire system. In general, the shape of the curves follows the same behaviour as for the case of the state threshold of 50%, where small increments of flow influence the mean value of information. However, the differences between the curves with different celerity values are less significant than for the cases with thresholds of 50% and 80%.

A very important feature revealed by Figure 6-15, Figure 6-16 and Figure 6-17 is that the mean value of information drops continuously as the state threshold increases. This is because the definitions of the state of the system given in Table 6-8 and Table 6-9 imply that only the excess of flow is of interest in defining a flood, and therefore whether to release a warning or not. Therefore, if the flood threshold is defined as percentile 50 of
the historic flow at every computational point, then the current state of the system can be deduced with more confidence than for the case of having a flood threshold of 80%. Alternatively, a threshold of 80% implies that 80% of the times the river is going to exceed the percentile 20% of the flow at every computational point. Naturally, a very low threshold indicates that the river is flooded frequently everywhere so that the state of the system is reasonably known and therefore the need to site monitors for this case is reduced. Conversely, it is more difficult to know if the river is going to exceed a threshold that is rarely reached and in this case the monitors will provide a larger value of information.

In practice, the state definition threshold should be replaced by the real critical flood levels in order to locate the monitors properly.

**Location of monitors**

The optimisation problem for the location of one, two and three monitors posed in the expressions (6-3), (6-4) and (6-5), is solved for celerity values between 0.5m/s and 3m/s. Additionally, three different state thresholds (80%, 50% and 20%) are selected in order to take into account different monitoring objectives. For example, the first threshold is for selecting the most appropriate locations for monitoring extreme events; the second is for selecting the locations that provide high quality messages about the “normal” state of the river (when the river exceeds the historic mean discharge at any point); lastly, the third threshold is to identify the monitor locations for low-flow monitoring. The results are summarized in Figure 6-18, Figure 6-19 and Figure 6-20.
<table>
<thead>
<tr>
<th>Celerity (m/s)</th>
<th>0.5</th>
<th>1.0</th>
<th>1.5</th>
<th>2.0</th>
<th>2.5</th>
<th>3.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>One monitor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two monitors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Three monitors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 6-18. Results for one, two and three monitor locations for different celerity values and 80% state threshold definition
Figure 6-19. Results for one, two and three monitor locations for different celerity values and 50% state threshold definition
Figure 6-20. Results for one, two and three monitor locations for different celerity values and 20% state threshold definition
From Figure 6-18 it can be concluded that for one monitor the point that provides the best quality of the message about the state of the entire system moves upstream as the celerity increases. However, the monitor is always located in the zone between the point of discharge of the Lebrija River and the wetland $W_2$, described previously. For a celerity value of 1 m/s a better distribution of the VOI is found, that is, more points in the river are described with one monitor. Possibly this is the physical celerity value for the Magdalena River. However, it is observed that only one monitor is insufficient to get messages about the state of the system at the wetlands and at the downstream part of the river.

From Figure 6-15 the location of two monitors for different celerity values appears sensible. Once again, the zone between the discharge of the Lebrija River and the wetland $W_2$ is selected because it has the highest $VOI_i$ value. The second region to place a monitor is downstream after the wetland zone, supporting in some way the job of the first monitor.

The result for three monitors can also be anticipated. It can be summarised as placing one monitor before, within and after the wetland zone. Naturally, the second monitor is placed near the discharge point from the Cauca River. When locating two and three monitors the effect of the celerity is not significant.

As mentioned above, the effect of the threshold reduction on the value of information takes place as the threshold decreases. This can be also observed in the Figure 6-18, Figure 6-19 and Figure 6-20. From these figures it can also be concluded that the monitor locations do not suffer major changes when using different celerity values, because they are determined largely by the three well-defined VOI zones in which the river is divided: the upstream reach, which can be monitored by at a single location downstream; the wetland zone, where important discharge fluctuations take place for the wetland-river connections and due to the discharge from the Cauca River; and the downstream reach, whose state can be monitored by a single point upstream.

6.6.3 Pijnacker region, The Netherlands

The VOI-based method for monitor location is also applied to the case study described in Chapter 3. For reasons related to computational effort, this section includes two different experiments: one with simplified inputs for the complete Pijnacker water system, and one with more complex inputs in a smaller system.

**Entire Pijnacker polder system**

For the first experiment, the same simplified inputs as for the canal-pump case are used and therefore the states, actions and messages are as shown in Table 6-5. Also, the consequence matrix of Table 6-4 is adopted. However, as the flood level information is available this was used for the experiment below.

The mean Value of Information for the entire Pijnacker polder system was estimated with Eq. (6-2), obtaining the VOI map shown in Figure 6-21.
It can be observed that importantly a percentage of the points do not provide value in terms of describing the state of the entire system. Also, the VOI is dependent on the location of pumps and weirs. It is interesting that most of these zero-VOI points are located at the most elevated parts of the system.

The location of one and of two monitors was determined and Figure 6-22 was prepared using the procedure described in Section 6.5,

It is noted that the solution of Eq. (6-3) for locating one monitor (Figure 6-22-a) yields two different results, while the solution of Eq. (6-4) for locating two monitors (Figure 6-22-b) yields seven separate results. One of the solutions is shown in both figures.
It is interesting that for the case of two monitors, the location obtained for the one monitor analysis is also selected and that after the placement of two monitors, an important number of points were still not covered from a VOI point of view. This means that in this water system, where the water level has a number of discontinuities due to weirs and pumps, it is too ambitious to pretend that the state of the entire system can be described with only two points. Although the procedure to solve Eq. (6-5) for the location of three monitors is not possible because of the computational resources needed for the number of calculation points under consideration, the division of the system into several continuous subsystems can be a way to overcome this issue. For this reason, the experiment was repeated for a subsystem, a simplification that allows more detailed inputs to be used, as explained below.

**Selected subsystem within Pijnacker polder system**
The selected subsystem is located at the North-East part of the Pijnacker water system. The water level in this area is controlled by four weirs and three pumps. The land is used mainly for pasture, but glasshouses and urban developments also exist. This subsystem was modelled with 65 calculation points (see Figure 6-23).

The availability of the model data allows three new features in the procedure for locating the monitors to be introduced. First, more than two possible states of the system, the definition of different land uses (urban, glasshouse and pasture), each one with its own damage function (consequence matrix) and the definition of experience-based consequence matrices are defined. These three aspects, discussed with staff members of the Delfland Waterboard in November 2009, are summarized in Table 6-10 and Figure 6-24, where four states, namely severe flood, flood, normal and drought are considered. Note that the consequences shown in Table 6-10, however, do not represent monetary
values but are relative costs, (the norm is 1000 units, being the reference for the worse-case scenario of doing nothing when a severe flood is present).

Table 6-10. Table of consequences $C_{as}$ for different land uses for the Pijnacker region

<table>
<thead>
<tr>
<th>$C_{as}$</th>
<th>$a_1$ Pump1 On</th>
<th>$a_2$ Pump1 Off</th>
<th>$a_1$ Pump2 On</th>
<th>$a_2$ Pump2 Off</th>
<th>$a_1$ Pump3 On</th>
<th>$a_2$ Pump3 Off</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1: Severe Flood</td>
<td>500</td>
<td>1000</td>
<td>250</td>
<td>500</td>
<td>50</td>
<td>100</td>
</tr>
<tr>
<td>S2: Flood</td>
<td>100</td>
<td>200</td>
<td>50</td>
<td>100</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>S3: Normal</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S4: Drought</td>
<td>20</td>
<td>10</td>
<td>10</td>
<td>5</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

Figure 6-24. Definition of the possible states, land uses and damage function (consequences)
Under these conditions, there are 32 possible situations that may happen in at each area of the water system. These situations are summarized in the Table 6-11 below.

<table>
<thead>
<tr>
<th>Situation</th>
<th>Message at $x$</th>
<th>State at $y$</th>
<th>State of the pump affecting $y$</th>
<th>Combined State ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>“Severe Flood”</td>
<td>S1: Severe Flood</td>
<td>On</td>
<td>SF-ON</td>
</tr>
<tr>
<td>2</td>
<td>“Severe Flood”</td>
<td>S2: Flood</td>
<td>On</td>
<td>F-ON</td>
</tr>
<tr>
<td>3</td>
<td>“Severe Flood”</td>
<td>S3: Normal</td>
<td>On</td>
<td>N-ON</td>
</tr>
<tr>
<td>4</td>
<td>“Severe Flood”</td>
<td>S4: Drought</td>
<td>On</td>
<td>D-ON</td>
</tr>
<tr>
<td>5</td>
<td>“Severe Flood”</td>
<td>S1: Severe Flood</td>
<td>Off</td>
<td>SF-OFF</td>
</tr>
<tr>
<td>6</td>
<td>“Severe Flood”</td>
<td>S2: Flood</td>
<td>Off</td>
<td>F-OFF</td>
</tr>
<tr>
<td>7</td>
<td>“Severe Flood”</td>
<td>S3: Normal</td>
<td>Off</td>
<td>N-OFF</td>
</tr>
<tr>
<td>8</td>
<td>“Severe Flood”</td>
<td>S4: Drought</td>
<td>Off</td>
<td>D-OFF</td>
</tr>
<tr>
<td>9</td>
<td>“Flood”</td>
<td>S1: Severe Flood</td>
<td>On</td>
<td>SF-ON</td>
</tr>
<tr>
<td>10</td>
<td>“Flood”</td>
<td>S2: Flood</td>
<td>On</td>
<td>F-ON</td>
</tr>
<tr>
<td>11</td>
<td>“Flood”</td>
<td>S3: Normal</td>
<td>On</td>
<td>N-ON</td>
</tr>
<tr>
<td>12</td>
<td>“Flood”</td>
<td>S4: Drought</td>
<td>On</td>
<td>D-ON</td>
</tr>
<tr>
<td>13</td>
<td>“Flood”</td>
<td>S1: Severe Flood</td>
<td>Off</td>
<td>SF-OFF</td>
</tr>
<tr>
<td>14</td>
<td>“Flood”</td>
<td>S2: Flood</td>
<td>Off</td>
<td>F-OFF</td>
</tr>
<tr>
<td>15</td>
<td>“Flood”</td>
<td>S3: Normal</td>
<td>Off</td>
<td>N-OFF</td>
</tr>
<tr>
<td>16</td>
<td>“Flood”</td>
<td>S4: Drought</td>
<td>Off</td>
<td>D-OFF</td>
</tr>
<tr>
<td>17</td>
<td>“Normal”</td>
<td>S1: Severe Flood</td>
<td>On</td>
<td>SF-ON</td>
</tr>
<tr>
<td>18</td>
<td>“Normal”</td>
<td>S2: Flood</td>
<td>On</td>
<td>F-ON</td>
</tr>
<tr>
<td>19</td>
<td>“Normal”</td>
<td>S3: Normal</td>
<td>On</td>
<td>N-ON</td>
</tr>
<tr>
<td>20</td>
<td>“Normal”</td>
<td>S4: Drought</td>
<td>On</td>
<td>D-ON</td>
</tr>
<tr>
<td>21</td>
<td>“Normal”</td>
<td>S1: Severe Flood</td>
<td>Off</td>
<td>SF-OFF</td>
</tr>
<tr>
<td>22</td>
<td>“Normal”</td>
<td>S2: Flood</td>
<td>Off</td>
<td>F-OFF</td>
</tr>
<tr>
<td>23</td>
<td>“Normal”</td>
<td>S3: Normal</td>
<td>Off</td>
<td>N-OFF</td>
</tr>
<tr>
<td>24</td>
<td>“Normal”</td>
<td>S4: Drought</td>
<td>Off</td>
<td>D-OFF</td>
</tr>
<tr>
<td>25</td>
<td>“Drought”</td>
<td>S1: Severe Flood</td>
<td>On</td>
<td>SF-ON</td>
</tr>
<tr>
<td>26</td>
<td>“Drought”</td>
<td>S2: Flood</td>
<td>On</td>
<td>F-ON</td>
</tr>
<tr>
<td>27</td>
<td>“Drought”</td>
<td>S3: Normal</td>
<td>On</td>
<td>N-ON</td>
</tr>
<tr>
<td>28</td>
<td>“Drought”</td>
<td>S4: Drought</td>
<td>On</td>
<td>D-ON</td>
</tr>
<tr>
<td>29</td>
<td>“Drought”</td>
<td>S1: Severe Flood</td>
<td>Off</td>
<td>SF-OFF</td>
</tr>
<tr>
<td>30</td>
<td>“Drought”</td>
<td>S2: Flood</td>
<td>Off</td>
<td>F-OFF</td>
</tr>
<tr>
<td>31</td>
<td>“Drought”</td>
<td>S3: Normal</td>
<td>Off</td>
<td>N-OFF</td>
</tr>
<tr>
<td>32</td>
<td>“Drought”</td>
<td>S4: Drought</td>
<td>Off</td>
<td>D-OFF</td>
</tr>
</tbody>
</table>

Therefore, there are two associated conditional matrices $q_{m,s}$ that depend on the current status of the pump downstream the point under consideration, each one with 16 elements. For the pump-on status, the $q_{m,s}$ matrix is presented in Table 6-12, whereas the $q_{m,s}$ matrix for the pump-off status is presented in Table 6-13.
### Table 6-12. $q_{m,s}$ matrix for the on-status of the pump downstream $y$ according to the situations of Table 6-11

<table>
<thead>
<tr>
<th>$q_{m,s}$</th>
<th>m1: “Severe Flood” (at x)</th>
<th>m2: “Flood” (at x)</th>
<th>m3: “Normal” (at x)</th>
<th>m4: “Danger” (at x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>s1 Severe Flood (at y), pump is on</td>
<td>Number of occurrences of situation 1 / Number of SF-ON</td>
<td>Number of occurrences of situation 9 / Number of SF-ON</td>
<td>Number of occurrences of situation 17 / Number of SF-ON</td>
<td>Number of occurrences of situation 25 / Number of SF-ON</td>
</tr>
<tr>
<td>s2 Flood (at y), pump is on</td>
<td>Number of occurrences of situation 2 / Number of F-ON</td>
<td>Number of occurrences of situation 10 / Number of F-ON</td>
<td>Number of occurrences of situation 18 / Number of F-ON</td>
<td>Number of occurrences of situation 26 / Number of F-ON</td>
</tr>
<tr>
<td>s3 Normal (at y), pump is on</td>
<td>Number of occurrences of situation 3 / Number of N-ON</td>
<td>Number of occurrences of situation 11 / Number of F-ON</td>
<td>Number of occurrences of situation 19 / Number of F-ON</td>
<td>Number of occurrences of situation 27 / Number of F-ON</td>
</tr>
<tr>
<td>s4 Drought (at y), pump is on</td>
<td>Number of occurrences of situation 4 / Number of D-ON</td>
<td>Number of occurrences of situation 12 / Number of D-ON</td>
<td>Number of occurrences of situation 20 / Number of D-ON</td>
<td>Number of occurrences of situation 28 / Number of D-ON</td>
</tr>
</tbody>
</table>

### Table 6-13. $q_{m,s}$ matrix for the off-status of the pump downstream $y$ according to the situations of Table 6-11

<table>
<thead>
<tr>
<th>$q_{m,s}$</th>
<th>m1: “Severe Flood” (at x)</th>
<th>m2: “Flood” (at x)</th>
<th>m3: “Normal” (at x)</th>
<th>m4: “Danger” (at x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>s1 Severe Flood (at y), pump is off</td>
<td>Number of occurrences of situation 5 / Number of SF-OFF</td>
<td>Number of occurrences of situation 13 / Number of SF-OFF</td>
<td>Number of occurrences of situation 21 / Number of SF-OFF</td>
<td>Number of occurrences of situation 29 / Number of SF-OFF</td>
</tr>
<tr>
<td>s2 Flood (at y), pump is off</td>
<td>Number of occurrences of situation 6 / Number of F-OFF</td>
<td>Number of occurrences of situation 14 / Number of F-OFF</td>
<td>Number of occurrences of situation 22 / Number of F-OFF</td>
<td>Number of occurrences of situation 30 / Number of F-OFF</td>
</tr>
<tr>
<td>s3 Normal (at y), pump is off</td>
<td>Number of occurrences of situation 7 / Number of N-OFF</td>
<td>Number of occurrences of situation 15 / Number of F-OFF</td>
<td>Number of occurrences of situation 23 / Number of F-OFF</td>
<td>Number of occurrences of situation 31 / Number of F-OFF</td>
</tr>
<tr>
<td>s4 Drought (at y), pump is off</td>
<td>Number of occurrences of situation 8 / Number of D-OFF</td>
<td>Number of occurrences of situation 16 / Number of D-OFF</td>
<td>Number of occurrences of situation 24 / Number of D-OFF</td>
<td>Number of occurrences of situation 32 / Number of D-OFF</td>
</tr>
</tbody>
</table>
Results
Following Table 6-1 to estimate the prior beliefs, the mean value of information $VOL_x$ is obtained for the case of considering all the records, and also considering only the records occurring when the associated pump station (pump downstream the current point $y$) is ON and OFF. Figure 6-25 presents the results for $VOL_x$ under these three scenarios. The most important observation, apart from the fact that the three maps differ only in few points, is that the maximum $VOL_x$ is 0.5 units, even though the maximum damage possible is 1000 units, when severe flood occurs in urban areas (see Table 6-10). This is because the records are mainly in the normal range at all points.

This implies that monitors will not be so valuable in providing messages about the state of the entire subsystem. That is, there is such a high confidence in the state of the system, that it is not worth placing monitors. Nevertheless, for the sake of testing the approach for monitor location, the expressions (6-3), (6-4) and (6-5) are solved. Due to the fact that many points have equal value of information, several solutions were obtained (43 solutions for one monitor, 303 for two monitors and 343 for three monitors for the pump-related data sets). The maps in Figure 6-26 show one of the possible solutions for each case.

These results confirm that even though the number of monitors increases, the value of information about the state of the entire subsystem remains low. However, it is clear that the monitors tend to concentrate mainly on the urban zone, where the damage function is the highest.

6.7 Conclusions

The VOI method for locating monitors optimizes VOI such that their messages can provide information about the state of the entire water system with minimum redundancy between the monitors.
The VOI approach to locating monitors is very flexible, because it is applicable to any type of water system, for any water variable and any set of statuses that may occur in the system. The use of models for data generation is interesting because it permits the analysis of a dense set of points from which those with the highest VOI can be selected.

There are two main difficulties when applying the method. First, there is the need for a good definition of consequences or damage functions. Second, a convenient definition of the states of the system that are to be monitored is required. The first difficulty concerns the amount of data required to define the damage functions, while the second involves the clear definition of the objectives of the monitoring network. Nevertheless, it has been shown that both difficulties can be resolved by analysing different scenarios.

One disadvantage with the method is the huge computational effort needed to solve the monitor location problem for more than three monitors. A workaround consists of dividing the water system into several subsystems and solving for each subsystem separately.

It is interesting that in the case of the Magdalena River the VOI is sensitive to the discharges from the tributaries and in the case of Pijnacker to the location of hydraulic structures. For the case of natural streams such as the Magdalena River, the inclusion of the dynamics of the kinematic wave allows for a better location of monitors, depending on the objective of the monitoring network.
Chapter 7
Public data collection and assessment of model reliability

In the previous chapters, different methods for placing monitor devices were developed and tested in the polder system of Pijnacker, described in Chapter 3 and in the river Magdalena, described in Chapter 4. In Chapter 5 the use of Information Theory concepts was explored in order to maximise the information content about the particular water system. A complementary method that takes into account the subjective beliefs of a decision maker in getting information is presented in Chapter 6 with the maximisation of the Value of Information. The present chapter explores other methodologies to obtain a reliable idea of the state of a water system through public data collection and model validation, which, in turn, looks for an incremental increase in the information content of our water systems.

Two main topics are distinguished in this chapter. First, the use of public participation in information collection, particularly by the use of mobile phones; second, the use of the data collected by this method to improve the reliability of the models for decision making. The first part describes the whole cycle, from conception to results, of an experiment called MoMoX (Mobile Monitoring Experiment), which was carried out during the first half of 2010 in the region of Pijnacker. The second part describes the methods by which the data collected in the first part can be used to improve models. Before entering into the details, the introduction describes the main motivations of the methods.

7.1 Introduction

Mobile phones are devices that have evolved significantly during the last years. More than just phones, these machines can be considered to be small computers that can offer a variety of features in addition to their basic communication possibilities. However, it is not only the technology behind them that makes them powerful, but also the social consequences derived from the fact that almost every person has one in his/her pocket.

Recently, several researches have explored different uses of mobile phones in water-related problems. On the one hand, mobile technologies can be used as a tool for spreading information, as demonstrated, for example, by Alfonso (2006), who presented a
methodology to use mobile phones to inform people about the proper times to consume drinking water in an intermittent water distribution systems, and Naz (2006), who used wireless technologies to spread early warning messages of flooding in Dakha. On the other hand, mobile phones are also a tool for public participation in monitoring, as suggested by Silva (2008), who works with children to create a multisensory geographical information system using mobile phones in learning and participatory contexts, and Gouveia et al. (2004) who overcame typical problems of voluntary data collection by promoting the use of ICT.

The motivation of the research presented in this chapter comes from a consideration of two aspects concerning data collection in extreme events. First, there is the problem of validating the results of rainfall-runoff models during such events when typical monitoring devices fail to record and transmit data. This situation leads to the incompleteness of the data records especially at the time of the peak of the event, which is actually the moment when a reliable model is needed to define the real extent of the flooding and to generate reliable information for decision making.

The second aspect is the fact that mobile phones are not simply communication devices but small computers that are part of a network, can transfer information in various formats, can accompany a human being, work at anytime and can connect to the Internet. Most importantly, people know how to use them and almost everybody has one. The implications for monitoring are clear: measurements can be sent and analyzed immediately, taken at any accessible place and at any time, and for as many times as required. Additionally, minimal training is needed, their use by human observers is flexible, and there are no maintenance costs and vandalism problems. The data that can be collected include water levels, rainfall, discharges (given velocity readings as well as water levels), the operational status of hydraulic structures such as pumps and gates, flood reports, failures, obstructions, etc.

This chapter is divided into two sections: the first addresses the introduction of Mobile Monitoring Experiment (MoMoX), an experiment using members of the public to collect water level data in a polder in The Netherlands; the second describes a methodology to use the collected data in an assessment of model reliability.

7.2 Public participation in data collection

In the Chapter 2 the related literature review was presented, where a number of projects with positive experiences in public participation are mentioned. In the following section the experience about letting people read and send water level data through mobile phones is described. Findings and limitations are mentioned.

7.2.1 Mobile Monitoring Experiment (MoMoX)

MoMoX stands for Mobile Monitoring Experiment, a new approach to collecting field data for water management purposes, which was held in the city of Pijnacker and nearby areas (see Chapter 3). The experiment was divided into three stages (see Table 7-1). The
first stage consisted of two small, pilot experiments that took place during February 2010, in order to test the technological platform and correct possible errors, one of them being held directly in the field. The second stage, with 15 participants and no rainfall event, was held during the 22nd of May, 2010 in the field, in order to evaluate the possible errors coming from people with no experience in water level gauge reading and the performance of the website. The third stage involved people living or working nearby the water level gauges, so that they are able to send messages during or after real rainfall events during the month of July 2010.

<table>
<thead>
<tr>
<th>Stage</th>
<th>Description</th>
<th>Duration</th>
<th>Number of Participants</th>
<th>Type of participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a) Pilot test field</td>
<td>2h</td>
<td>4</td>
<td>Colleagues</td>
</tr>
<tr>
<td></td>
<td>b) Pilot test lecture room</td>
<td>2h</td>
<td>9</td>
<td>Students</td>
</tr>
<tr>
<td>2</td>
<td>Real field test</td>
<td>1 day</td>
<td>15</td>
<td>Students + colleagues</td>
</tr>
<tr>
<td>3</td>
<td>Real field test</td>
<td>1 month</td>
<td>2</td>
<td>Residents</td>
</tr>
</tbody>
</table>

The platform for the experiment consists of a website with details of MoMoX, which include the location of the gauges in the region, instructions to read water level gauges and instructions to send an SMS with the water level information. Every participant can register either through filling in an online form or by sending an SMS with the nickname of his/her choice. The experiment procedure has the following steps, noting that all the gauges in the region were previously labelled by staff members of Delfland Waterboard.

- An SMS with the expected rainfall amount for the coming days is prepared by Hoogheemraadschap van Delfland, the local water authority, and sent to the field operators and to MoMoX.
- An SMS, indicating the start date and time for the experiment and the phone number where the messages are to be received, is sent to the registered participants.
- The participants start visiting as many water level gauges as possible, read them and send messages in a predefined format.
- The SMS data are received and validated automatically on the MoMoX server.
- The data is displayed on the website so that the participants can see their records on a map display. A list of the top-10 contributors is continuously displayed on the website, to encourage participation.
- The water level information can be immediately accessed by analysts in the office. Provided the experts have an Internet connection, they can also check the water level behavior at different places and take proper decisions.
Figure 7-1 presents these steps graphically, in the form used in a flyer (produced in both Dutch and English) to advertise the experiment and to encourage people to participate.

7.2.2 Technology used

The technology and processes used in MoMoX are schematized in Figure 7-2. The process begins with the generation of a mobile-originated (MO) message from a mobile user. The transmission of SMS is done through the local cellular network by an API gateway provided by Clickatell, which makes a post with the contents and the metadata of the SMS to a predefined target address in the web. This post is then captured by a PHP code located in the mentioned target address, which analyses it and extracts, among other data, the source of the message (mobile number), the text message and its reception time.

Additionally, the PHP code performs the following tasks:
- Extract the gauge identification number and the reading value from the text message.
- Using the gauge identification number, the high and low water level limits are checked. If the reading value is outside this range, a feedback message is sent back to the user, to let him/her know that there may be a mistake in the reading.
Chapter 7 - Public data collection and assessment of model reliability

- The reading value is converted to absolute elevation using the NAP reference. This information is of interest for the operators of the system, since they can check whether the target water levels are met with the current control strategies.
- The incoming data is stored in a server database, on which SQL-like commands are performed to draw the stage graphs at each measured point using Google Visualization API. The map showing the gauge locations is Google Maps-based. Every time the user clicks on a gauge, the SQL statement is activated, the visualization API is called and the graph is drawn (see Figure 7-3).

![Diagram of technology behind MoMoX](image)

**Figure 7-2. Technology behind MoMoX**

![MoMoX website showing gauge 8 info, and the current water level graph.](image)

**Figure 7-3. MoMoX website showing gauge 8 info, and the current water level graph.**
As the database is being fed in real-time, the graphs are immediately available for checking. An important way of encouraging participants to send messages is by identifying the contributor (nickname) of each single point in the graph. This, in addition, provides an indirect way of identifying the people who send wrong data.

### 7.2.3 Communication campaign

As the central part of the experiment is the use of the general public to collect data, the experiment has a strong social component, in which the communication processes key in its success.

Due to the trial-nature of the first stage, a major communication process was not really required. The pilot test in the field demanded only little time from four colleagues at UNESCO-IHE. For the pilot test in a lecture room, a presentation of the current research given to the MSc students of Hydroinformatics at UNESCO-IHE was utilized to meet potential participants. During the presentation, a simulation of the real experiment was carried out. As expected, their registration process was straightforward.

For the second stage, however, a stronger communication campaign was needed, in order to collect as many participants as possible for a real data collection campaign. The strategy included, on the one hand, disseminating the message throughout the whole UNESCO-IHE, by email, by personal approach and by letting a video run in public screens at the entrance of the institute; on the other hand, handing out letters to the residents living in the area, preferably nearby the gauges to be read. A total of 85 letters, in Dutch language, were submitted. As a result, 28 participants registered through the online form in the MoMoX website, 2 of them being residents of the area. On the day of the experiment, however, 15 people were actually in the field, 12 of them being UNESCO-IHE students, 1 staff member of the same institute and 2 people with other affiliations. Unfortunately, the inhabitants that registered did not send any SMSs during the experiment.

The first two stages of the experiment exposed the need for stronger communication campaigns to involve participants. For this reason, the third stage consisted of 25 house-by-house personal visits to explain the experiment and to hand out flyers in Dutch. In order to attract their attention, 15 umbrellas were given away to enthusiastic participants. However, approximately half of the people contacted was enthusiastic enough (especially those with scientific background), while the other half were either not interested, not willing to listen the proposal or too busy. Key communication players, such as the local newspaper, schools and the association “Vereniging voor Natuur- en Milieubescherming Pijnacker”, an association for nature conservation in the Pijnacker region, were also contacted. However, as a result of this campaign, only five residents actually registered.

### 7.3 Results of the experiment

The results obtained for the three stages mentioned previously, namely the pilot tests, the experiment of 22nd of May 2010 and the one-month experiment are presented below.
7.3.1 Findings of the pilot tests

Two main pilot tests were carried out during February 2010. The first one, executed by four participants, consisted of sending messages from four different locations in the field.

During the first test it was found necessary to provide information on the scales of the gauges, since some of them are in meters, others in decimeters and still others in centimeters, and therefore very different values were being reported for the same site (Figure 7-4).

In addition, a command to send feedback messages to the participants was found to be needed in case the values were rejected during the validation procedure. In this way, the participants have the opportunity to resend the value after an explanation of the possible sources of the mistake. Finally, it was necessary to give the option to provide specific information for a given gauge at the user’s request, such as longitude, latitude, scale and picture (the latter for those with the option of receiving MMS).

The second pilot test was carried out during a Hydroinformatics’ lecture at UNESCO-IHE, in which 9 participants with Internet access were involved. The objective of this test was to check how the platform would handle several messages coming simultaneously from different mobile operators and how the website would perform when simultaneous clients access the maps and the graphs. For this purpose, a visit to the field was simulated by providing the message the people would send.

Two main issues were found during the second pilot test:

- Mistakes during the mobile numbers registration, especially regarding the format of the mobile numbers in international format, caused the system not to be able to receive the inputs from two participants.
- Participants subscribed to the mobile carriers KPN Mobiel, Orange Nederland, T-Mobile Nederland BV, Tele2, Telfort (O2) and Vodafone (Libertel), were able to send messages, while participants with other carriers were not. Unfortunately, this
issue is exclusively dependent of the SMS gateway provider (Clickatell), as other carriers are not included currently in their service. Similarly, sending SMS from web-based services was not possible.

In spite of these difficulties, the test demonstrated that the platform is a robust one, and that the graphs and tables are successfully updated in few seconds after the SMSs are sent.

7.3.2 Findings of the experiment of May 22

The second stage of the experiment consisted of testing the platform from the field, during a one-day activity with several participants. The purpose of this experiment was to assess the reliability of the data sent by each participant after reading the water level gauges by themselves. For this reason, validation-related errors were corrected offline in order to exclude them from the reading errors, which is the issue of interest. It must be mentioned that the experiment was carried out in a warm, dry day, and that there was no evidence of pump operation, so the water levels remained constant all day long.

The validation-related problems found during this experiment are listed as follows:

- One of the participants used comma (,) as decimal separator, instead of period (.), generating false data. The resulting gauge value appeared to be the digit before the comma, and the gauge ID the digits after the comma.
- A problem related to the negative sign was found when a participant sent a datum with the negative sign separated by a space. As the space field is used as a separator, the system interpreted this as an error and the datum was not registered.
- One of the participants sent a gauge value, but forgot to add the gauge ID. The automatic feedback SMS warned this person and the message was resent.
- Although the validation process includes the check of gauge limits, other procedures for avoiding scaling mistakes were found to be needed. For instance, (see Figure 7-5), one of the participants sent the value 0.2 instead of 2 for the gauge 8, due to an error of appreciation of the gauge scale; yet, the value of 0.2 appeared to be within the valid scale range. The same situation occurred in the gauge 3, where a participant sent -0.1 instead of -0.01.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Gauge 3</td>
<td>0</td>
<td>-0.02</td>
<td>-0.06</td>
<td>-0.08</td>
<td>-0.12</td>
<td>0</td>
<td>-0.02</td>
<td>-0.06</td>
<td>-0.08</td>
</tr>
<tr>
<td>Gauge 8</td>
<td>0</td>
<td>-0.5</td>
<td>-1</td>
<td>-1.5</td>
<td>-2</td>
<td>-2.5</td>
<td>-3</td>
<td>-3.5</td>
<td>-3.5</td>
</tr>
</tbody>
</table>

Figure 7-5. Examples of validation errors related to gauge scale
However, the check limit procedure avoided the registry of a wrong datum sent by different participants in 5 opportunities at the gauge 19 and one at the gauge 16. Note that these gauges have scales in meters (e.g., Figure 7-4c), so this situation points out a particular difficulty.

On the other hand, problems related to the SMS generation were also found and are listed below:

Some participants could not find the negative sign in their mobile phone text options for SMS when redacting the message. Therefore, they send either a positive value (see e.g., Figure 7-6) or simply did not send anything.

![Gauge 13](image1.png)

**Figure 7-6. Example of error related to datum without negative sign.**

However, the very opposite case occurred in the gauge 6, where a participant sent a negative value when the true value was a positive one. As the datum is within the limits of the gauge scale, it was accepted by the system (see Figure 7-7). This might happen because of the lack of concentration, maybe due to tiredness. In fact, physical effort is needed to access this gauge, as it is located below a bridge in the middle of the long pasture grass, which also make it difficult to read the gauge comfortably.

![Gauge 6](image2.png)

**Figure 7-7. Example of error related to adding an unnecessary negative sign.**
As previously mentioned, the gauges with scale in meters were found to be the most difficult to read by the participants, probably because the addition of decimal places is a source of confusion. However other reasons can be that both gauges (16 and 19) have small figures, and their readings must be done from a considerable distance.

As expected, random errors related to the appreciation of the gauge scales were found. Some of the resulting water level graphs shown in Figure 7-8 can be used to analyse these errors.

It can be observed that the random errors are the order of 2 cm. The same error is obtained if the outliers identified before (shown from Figure 7-5 to Figure 7-7) are removed.

Figure 7-8. Example of random errors due to differences in appreciation

7.3.3 Findings of the experiments with residents

A total of 6 rainfall events occurred during July 2010 in the region of Pijnacker. For each of these events, a text message was sent to the participants, stating that a rainfall event was coming and that a SMS was expected from them.

During the whole month, two residents sent SMSs, namely ‘Rein’, who sent 6 SMSs from the gauge number 19 and ‘Eric’, who sent 4 from the gauge number 12 (see Figure 7-9)
Feedback from participants
Telephone interviews were held at the end of the experiment with those who registered. Unfortunately, three of them did not answer the phone call, which may imply that they simply were not available.

The two residents who took part in the experiment, explicitly stated that they joined it because for them it was a good idea, had a clear purpose and required low effort on their part. Additionally, both participants explained that the holiday period was not good for participating in the experiment.

The interviews made it clear that the flyers and the face-to-face visits were clear enough, so in principle these did not explain the low participation. However, one of the participants admitted his initial scepticism about the study, as he thought it was “a kind of call-game for which I had to pay a lot of money”. Also he stated that “it is disappointing that my neighbours did not participate because I could not compare or share my inputs, but nice that I knew when it was going to rain thanks to the SMS”.

Future research should explore the use of an extensive media campaign to encourage public participation, by reporting the successful results of this first experiment.

7.4 Assessing model errors with public’s data

The second section of this chapter describes how the data gathered by members of the public can be used for model validation. The proposed method is developed for a sub-region of the Pijnacker case study (see details in Chapter 3), in which the water levels for seven subareas with different elevations are controlled by means of four weirs and three pump stations.
Areas A1, A2, A5 and A6 are high level areas that discharge to A4 via the weirs 1, 2, 3 and 4 respectively. Area A3 needs the pump 2 to discharge to A4 because the latter is at a higher level. Next, all water collected in A4 is pumped to a higher level (A7) by means of the pump 2, to be finally discharged to the main storage canals by the pump 3.

Figure 7-10. Description of the area, absolute elevations and location of hydraulic structures

7.4.1 Description of the validation method

The validation method consists of two components:

a) The creation of a library of patterns in which a large number of models are generated with different inputs, including extreme events, in a Monte Carlo-like procedure.

b) The selection of the pattern that best describes the pattern depicted by the incoming public-generated data.

Generation of a library of patterns

In order to generate the library of patterns, the possible variables to be changed are defined. For the weirs, these variables are the crest level, discharge coefficient and length; for the pumps, the on and off levels and their capacities are considered; for the canals, the variables are the roughness coefficients and the dimensions of the cross sections; in order to vary the external inflows, a single event that is affected by a set of factors is considered.

The library of patterns is generated by changing all the variables within realistic value ranges. However, due to the huge number of model runs required, the number of variables was reduced by selecting the most uncertain variables. In this way, it is assumed that the (fixed) crest levels and the length of the weirs are known with reasonable accuracy. Similarly, the on/off levels of the pumps are considered to be well defined.
Additionally, the dimensions of the cross section are assumed to be constant since erosion and sedimentation processes are not significant in the area.

This means that four variables remain for selection: the weir discharge coefficients (to be separately changed for each of the four weir structures), the pump capacities (different for each pump), the roughness coefficient and the factor for the inflows. In consequence, a total of 1152 different model runs were generated and executed, and the corresponding water level time series at the available gauges $g_1$, $g_2$, $g_3$, $g_{17}$, $g_{18}$ and $g_{20}$ (see Figure 7-10) were stored in a library.

**Pattern selection using coming SMS data**

The second component of the method for model validation is the selection of the patterns according to the coming SMS data from the members of the public, as a way to select good model inputs and to improve the available (offline) model. Here a distinction between online (real-time) and offline can be made. In this section, only a method for offline models is presented.

The procedure to validate an offline model consists of selecting those patterns $P$ that best fit the SMS-based series according to, for instance, a linear least squares evaluation. If more than one pattern is selected, they equally represent the SMS data at the gauged points in the system and the choice of patterns may not be sensitive to some variables.

In order to improve the available model used for operation and decision support (subsequently called the zero-model), a comparison between the zero model and every pattern in $P$ is carried out, by comparing the variables used in each pattern. Then, the probability of “goodness” of the zero-model is defined as the number of times that the parameter is right divided by the number of total parameters. In the case that $P$ has more than one pattern, the redundant variables are removed from the set and the probability is assessed using the remaining variables.

**7.4.2 Results of the validation method**

The zero-model of the region was given. Although the initial idea was to use public-based SMS data, due to the low participation of the residents and therefore the few data collected, artificially generated SMS were used to demonstrate the procedure for model validation. The SMS-based data was generated using the same model with inputs that were not used for the construction of the library of patterns. The inputs of both models are shown in Table 7-2.

<table>
<thead>
<tr>
<th>Model</th>
<th>V1</th>
<th>V2</th>
<th>V3</th>
<th>V4</th>
<th>V5</th>
<th>V6</th>
<th>V7</th>
<th>V8</th>
<th>V9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero-model</td>
<td>1.0</td>
<td>0.07</td>
<td>0.1</td>
<td>0.5</td>
<td>1.5</td>
<td>2.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>SMS</td>
<td>1.3</td>
<td>0.04</td>
<td>0.1</td>
<td>0.5</td>
<td>1.5</td>
<td>1.0</td>
<td>2.0</td>
<td>1.0</td>
<td>2.0</td>
</tr>
</tbody>
</table>
From the prebuilt library of patterns, three patterns that fit the SMS data better were found; their characteristics are shown in the Table 7-3. It can be observed that the model output is not sensitive to the Manning coefficient and therefore this parameter can be neglected in the probability calculations that are detailed below.

<table>
<thead>
<tr>
<th>Model</th>
<th>V1</th>
<th>Inflow</th>
<th>V2</th>
<th>V3</th>
<th>V4</th>
<th>V5</th>
<th>V6</th>
<th>V7</th>
<th>V8</th>
<th>V9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pattern 1</td>
<td>1.0</td>
<td>0.01</td>
<td>0.1</td>
<td>0.5</td>
<td>0.5</td>
<td>2.0</td>
<td>1.0</td>
<td>2.0</td>
<td>2.0</td>
<td></td>
</tr>
<tr>
<td>Pattern 2</td>
<td>1.0</td>
<td>0.02</td>
<td>0.1</td>
<td>0.5</td>
<td>0.5</td>
<td>2.0</td>
<td>1.0</td>
<td>2.0</td>
<td>2.0</td>
<td></td>
</tr>
<tr>
<td>Pattern 3</td>
<td>1.0</td>
<td>0.07</td>
<td>0.1</td>
<td>0.5</td>
<td>0.5</td>
<td>2.0</td>
<td>1.0</td>
<td>2.0</td>
<td>2.0</td>
<td></td>
</tr>
</tbody>
</table>

A visual comparison of the zero-model, the SMS data and the best patterns retrieved, is presented in Figure 7-11 for the gauged points located in Figure 7-10.

It can be seen that the zero-model differs from the retrieved patterns in the capacity of pump 3 and in the discharge coefficients of the weirs 3 and 4. The probability of
goodness of the zero-model is therefore, \( p_g = \frac{6}{9-1} = \frac{3}{4} \); the remaining 25% of wrongness can be then reduced by adopting the values of V5, V8 and V9 for any of the retrieved patterns.

It must be noted that the patterns that best replicate the SMS data have been artificially generated using a different inflow factor from the one used in the Zero-model. This implies that the difference in the inflow factor is being compensated by the capacity of pump 3 and the discharge coefficients of weirs 3 and 4, and therefore the selected patterns may provide a wrong representation of what is happening in reality. However, once real, public-based SMS is available, the method will provide real insights into the source of errors in the Zero-model.

7.5 Conclusions

While traditional monitoring generally provides enough acceptable data to calibrate and validate models this is not the case during extreme events. In this chapter we demonstrate that the combination between public participation and mobile phones provides a promising way to deal with this problem.

It is necessary to generate a more comprehensive library of patterns in order to cover a wider range of possible inputs, including extreme events.

The validation of the SMS data coming from members of the public is a main issue that needs further research. For this, different approaches, such as the use of the improved model itself in a cyclic procedure, and the use of pattern recognition and image processing of mobile-originated pictures are worth being explored. Additionally, mechanisms to encourage public participation in order to have a denser data set would help to identify those SMS data that are incorrect.
Chapter 8
Conclusions and recommendations

In this thesis new methods of optimising monitoring networks using concepts of Information Theory, Value of Information and public participation have been investigated and tested in two case studies with distinct hydrologic, hydraulic, socioeconomic and political conditions. The conclusions and the recommendations of the research are presented in the sections below, following the same order in which the contributions were developed and applied.

8.1 Conclusions

8.1.1 General conclusions

It has been demonstrated that monitoring networks can be optimised by maximising the information content and information value, and in addition, that it is possible to configure a public-based monitoring network with mobile phones. As informed decision-making is key for adequate water management, the efforts presented to design and evaluate monitoring networks will lead to an enhancement of the performance of a water system.

Socioeconomic and political conditions drive, in practice, the design and the evolution of monitoring networks. In developing countries, lack of financial resources is the main reason for inadequate network density, so their efforts generally concentrate on optimising the use of the few monitors available. In contrast, in developed countries, efforts are generally addressed to reduce the size of the existing, dense monitoring networks.

8.1.2 Information theory for designing monitoring networks

Three new methodologies for optimising monitoring networks by coupling Information Theory concepts and models were successfully developed and applied, namely the Water Level Monitoring Design in Polders (WMP), the Multi Objective Optimisation Problem (MOOP) approach and a rank-based greedy algorithm. The following conclusions can be made:

The Information Theory capability of quantitatively measured information is a feature that is intensively exploited in this thesis for monitoring design. The insights it provides
about the distribution of the information content is invaluable for optimally place monitoring devices.

The distribution of the information content for a water system is driven by features that produce changes in the hydraulic conditions of the system. For the case of the polders of Pijnacker, the features that interfere with the distribution of the information content are the weirs and the pumps; for the case of the Magdalena River, it is the incidence of its wetlands and tributaries. However, this does not mean that a monitoring network configured by placing monitors at the hydraulic structures in the first case or at the tributaries in the second case will be optimal from an Information Theory perspective.

In all the experiments, the trade-off between the amount of information content provided by a set of monitors and the independency among them (mathematically represented by Joint Entropy and Total Correlation) is evident. However, a large Joint Entropy can be preferred over a small Total Correlation, because a low-dependent set of monitors is also the least informative. In such a case, therefore, it does not make sense to have a completely independent set of monitors if they do not provide enough information content individually.

The location with the highest information content in the system is usually the most dependent with regards to the remaining locations. This implies that once this point is selected, it is very difficult to find a second point that is informative and at the same time independent. This also explains why only a few monitors are needed in spite of having many possibilities (8 out of 1520 potential monitors for the Pijnacker region and 7 out of 181 for the case of the Magdalena River).

The sensitivity of any discretization-based criteria for the assessment of probability distributions is a well-known difficulty that greatly affects the estimation of entropy-related quantities. However, these effects are negligible for the methods determining monitor locations due to the relative nature of the developed methods.

8.1.3 Value of Information for designing monitoring networks

It was proved that monitoring networks can be designed by combining the aspects of VOI and the encapsulated knowledge that the modelling technology offers. The following conclusions can be drawn from this investigation:

The VOI-based approach to locate monitors is very flexible, because it allows for the analysis of any type of water system, for any water variable and for any set of states that may occur in the water system. The use of models for data generation is needed because it permits the analysis of a dense set of points from which those with the highest value are selected.

The VOI is sensitive to the discharges from the tributaries in the case of the Magdalena River, and to the location of hydraulic structures in the case of Pijnacker. This is because the definition of the states to be monitored in each case study is directly related to the behaviour of these elements.
The research makes clear that there is the need for a good definition of consequences or damage functions and for convenient definitions of the states of the system that are to be monitored. The difficulty in the first case is the amount of data required to define such damage functions, while for the second it is the clear definition of the objectives of the monitoring network. Nevertheless, both obstacles can be overcome by analysing different scenarios, in such a way that a variety of monitoring networks can be produced to replicate the states of the system that are required for water management.

The huge computational effort needed to solve the monitor location problem for more than three monitors can be resolved by dividing the water system into several subsystems and solving the problem for fewer monitors in each subsystem. In this case, it must be noted that the selected monitoring devices might not be the most valuable to describe the state of the system of the entire network, but only the state of the subsystem under consideration. Moreover, this workaround cannot guarantee the independency between monitor locations, and therefore subsequent analyses in this regard should be carried out.

For the case of natural streams such as the Magdalena River, the inclusion of the dynamics of the kinematic wave allows for a better location of monitors, depending on the objective of the monitoring network.

8.1.4 Public participation in data collection and model improvement

A public-based monitoring network for water-level data collection, characterised by the use of mobile phones, has been configured, tested, run and assessed. The main findings of this stage of the research are listed below.

The data sent by public SMS shows that random errors are of the order of 2 cm, once other errors such as negative signs and scale misleading are excluded by validation processes.

The combination between public participation and mobile phones provides a promising way to deal with the problem of collecting data for calibrating and validating models in the case of extreme events. The communication campaign to get people involved is, however, a major feature. For this reason, future research should explore the use of an extensive media campaign to encourage public participation, for example by radio or TV, by reporting the successful results of this first experiment.

From the feedback received from the residents, it can be concluded that short-horizon rainfall forecast information, for example 15min to 2hr in advance, is a possible successful business.

Although not officially part of this thesis, a parallel application of the concepts developed in this part of the research was carried out in the urban basin of the Tunjuelo River in Bogotá, Colombia, under the supervision of Ir. Carolina Rogéliz, following fruitful discussions. In this case, residents characterised as having low income, read and send raingauge data through SMSs that are evaluated to provide feedback about flood risk.
This positive experience confirms the benefit of public participation and mobile phones in data collection and corroborates the impact of this research in developing countries.

8.2 Recommendations

The fundamental concepts behind Information Theory and Value of Information proved to be complementary in the problem of designing and evaluating a monitoring network. It is expected that the developed methods can be combined into one, comprehensive method that takes into account both information content and decision-making aspects.

Certainly, this research opens up new possibilities for the application of Information Theory in water resources, in which information has been traditionally quantified in entropy units. With the inclusion of the Value of Information concept, the entropy-based methods developed in this thesis, as well as the ones developed in previous studies, can now consider measuring information in monetary terms. In this regard, the findings of this thesis can be used as the starting point to find the value of one unit of information.

Regarding the method for improving models with publicly collected data, it is necessary to generate a more comprehensive library of patterns as described in Section 7.4.1 in order to cover a wider range of possible inputs, including extreme events. Additionally, this method should be tested with real SMS coming from the public.

The use of pattern recognition and image processing of mobile-originated pictures are worth exploring, not only to gather information about the state of the system, but also to be adapted into a framework for validation of SMS data. Additionally, mechanisms to encourage public participation in order to have a denser data set would help to identify those SMS data that are incorrect.


Alvarado, M. "Desarrollo de proyecto piloto de navegacion satelital – SNS, entre Puerto Berrio (k783) y Regidor (k454)." *XXII Congreso Latinoamericano de Hidráulica*, Cuidad Guayana, Venezuela.


Optimisation of monitoring networks for water systems


Silva, M. I. G. "PLAN DE MANEJO DE LOS RECURSOS ICTIOLÓGICOS Y PESQUEROS EN EL RIO GRANDE DE LA MAGDALENA Y SUS ZONAS DE AMORTIGUACIÓN."


List of figures

Figure 1-1. One of the main challenges when designing a monitoring network .......... 6
Figure 1-2. Use of models for the design of monitoring networks (this thesis) .......... 8
Figure 1-3. Outline of the thesis .................................................................................. 10
Figure 2-1. First cyclic approach for monitoring planning........................................ 12
Figure 2-2. Classification of methods for design and evaluation of monitoring networks ......................................................................................................................... 14
Figure 2-3. Flowchart for VOI estimation ..................................................................... 23
Figure 3-1. Limits of the Delfland Water Board and in the province of Zuid Holland .... 28
Figure 3-2. Land uses, main water system components of Delfland region and location of the polders of Pijnacker. Adopted from Lobbrecht (1997) ....................... 29
Figure 3-3. Interest-weighting chart Delfland water system. Adopted from Lobbrecht (1997), p 183 .............................................................................................................. 29
Figure 3-4. Land use in the region of Pijnacker.............................................................. 30
Figure 3-5. Schematic profile of the polders of Pijnacker ............................................. 30
Figure 3-6. Composition of the polder system of Pijnacker and identification of pump stations ...................................................................................................................... 31
Figure 3-7. Canal network and target water levels in the polders of Pijnacker.............. 32
Table 3-1. Characteristics of the pump stations in the Pijnacker polders...................... 33
Figure 3-8. Location of the existing water level gauges in the Pijnacker polders .......... 34
Figure 3-9. Connection points of the hydrodynamic (HD) model and the rainfall-runoff (RR) model ............................................................................................................... 35
Figure 4-1. General location of the Magdalena River and its catchment ....................... 38
Table 4-1. Mean hydraulic slope by sectors, Magdalena River .................................... 38
Figure 4-2. Main tributaries and towns of the middle and low Magdalena River ........ 39
Figure 4-3. Mean discharges of the main tributaries of the middle and low sector of the Magdalena River ........................................................................................................ 40
Figure 4-4. Inner delta and wetlands of the Momposina depression ............................. 41
Figure 4-5. Age of the existing limnigraphic and limnimetric stations in the middle and low Magdalena River (based on the year 2010) ......................................................... 42
Figure 4-6: Available hydrologic data records of discharges (Q) and water levels (h) at river stations for 1995. .................................................................................. 46
Table 4-2. Number of days of 1995 with discharge and stage data and datum of gauges ......................................................................................................................... 46
Figure 4-7. Mean discharge of tributaries of the Magdalena River and mean discharges for the year 1995, used as model inputs ............................................................ 47
Figure 4-8. Example of a composite cross section near La Dorada - Puerto Salgar ...... 49
Figure 4-9. Assumed grouping of the wetland system for the model......................... 50
Table 4-3. Characteristics of the grouped wetlands..................................................... 51
Figure 4-10. Modelled and measured discharges at Regidor (a) and Calamar (b), first
check .............................................................................................................................. 52
Figure 4-11. Modelled and measured discharges at Regidor (a) and Calamar (b), second
check .............................................................................................................................. 52
Figure 4-12. Modelled and measured discharges at Regidor (a) and Calamar (b), final
result ............................................................................................................................. 53
Figure 4-13. Modelled and measured discharges at Santa Ana station, Mompox
branch (final result) ........................................................................................................ 53
Figure 4-14. Modelled and measured absolute water levels at Regidor station (final
result) ............................................................................................................................. 54
Figure 5-1 Rainfall event used in the hydrodynamic model............................................. 62
Figure 5-2. Example of original water level time-series and its quantized version, at a
point located downstream a pumping station................................................................. 63
Figure 5-3. Entropy map of Pijnacker region ................................................................... 63
Figure 5-4. Directional Information Transfer Index ($DIT_{XY}$) map for the point A (bits). 64
Figure 5-5. Step-by-step solution for the location of water level monitors using
$I(\mathbf{X};\mathbf{Y})$ as the dependency criteria. Entropy of the currently selected point is
shown at each step.......................................................................................................... 65
Figure 5-6. Step-by-step solution for the location of water level monitors using
$DIT_{XY}$. Entropy of the currently selected point is shown at each step (bits). .......... 66
Figure 5-7. Step-by-step solution for the location of water level monitors using
$DIT_{YX}$. .......................................................................................................................... 66
Figure 5-8. Location of water level monitors obtained by the WMP approach, using
$I(X;Y)$, $DIT_{XY}$ and $DIT_{YX}$ as pairwise dependency criteria................................. 67
Figure 5-9. Evolution of the values of Joint Entropy and Total Correlation as new
monitors are added to the solution set............................................................................ 68
Figure 5-10. Average percentage of common monitor locations comparing the
solution obtained for each value of $\alpha$ with all other calculated solutions
($\alpha=1,2,\ldots,10,15,20$) ............................................................................................. 70
Figure 5-11. Venn diagrams illustrating the proposed optimization problem. (A):
Information content of 10 variables and their common information; (B), (C) and
(D): possible solutions for the selection of three monitor locations (1), obtained by
maximizing joint entropy (2) and minimizing Total Correlation (3). ......................... 73
Figure 5-12. Definition of low-entropy points to be discarded from the search space
for the optimization, according to the relative frequency of the entropy of the points
in the Delfland system.................................................................................................... 75
Figure 5-13 Pareto-optimal set of solutions discriminated by EPM, approach a).
Extremes $X_a$ and $Y_a$ are indicated for further analysis. Results obtained with WMP
method (Alfonso et al. 2010b) are also indicated. .......................................................... 77
List of figures

Figure 5-14. Delfland water system with location of solutions for approach a) obtained at the extremes $X_a$ and $Y_a$ of the Pareto frontier of Figure 5-13. Solution for $S=EPM$ is also included. Scale represents the marginal entropy at each system point estimated with Eq. (2-1). ................................................................. 78

Figure 5-15. Pareto-optimal front, approach b), discriminated by the number of times the minimum distance $d_s$ is violated by the solution set. Extremes $X_b$ and $Y_b$ are indicated for further analysis. Results obtained with WMP method (Alfonso et al. 2010b) are also indicated. .............................................................................. 80

Figure 5-16. Delfland water system with location of solutions for approach b) obtained at the extremes $X_b$ and $Y_b$ of the Pareto frontier of Figure 5-15. Location of existing hydraulic structures is also included. Scale represents marginal entropy values at each system point (bits). ........................................................................ 82

Figure 5-17. Progress of information quantities as new monitors are added. Analysis for extremes $X_a$ and $Y_a$ of Figure 5-13 ........................................................................................................ 83

Figure 5-18. Progress of information quantities as new monitors are added for solution of approach a), for extremes $X_b$ and $Y_b$ of Figure 5-15. ............................................................... 84

Figure 5-19. Sensitivity of the maximum Joint Entropy (1) and Total Correlation (2) due to variations of the parameter $u$, discriminated by the number of new monitors in the solution. ........................................................................................................ 84

Figure 5-20. Sensitivity of the maximum Joint Entropy (left) and Total Correlation (right) due to variations of the parameter $q$, discriminated by the number of new monitors in the solution. ............................................................... 85

Figure 5-21. Flowchart rank-based greedy algorithm for Joint Entropy (a) and Total Correlation (b) ........................................................................................................ 88

Figure 5-22: Available hydrologic data records of discharges (Q) and water levels (h) at river stations for 1995. ........................................................................................................ 89

Figure 5-23. Entropy Map for $a=200$ m$^3$/s in bits (a) and mean discharge map for 1995 in m$^3$/s (b), for the Magdalena River ........................................................................................................ 91

Figure 5-24: Solutions for multiobjective optimization approach. Black dots form the best Pareto front obtained by selecting the best points of the 5 combinations (P, G). Points A, B, C and D are selected for further analysis for 9 decision variables. ..................... 92

Figure 5-25. Location of selected solutions A, B, C and D of Figure 5-24 for 9 decision variables ............................................................................................................. 93

Figure 5-26. Location of the most informative (and redundant) solution obtained for 6, 7, 8 and 9 monitors (the most right black dots of each Pareto front of Figure 5-24) ........................................................................................................ 94

Figure 5-27. Results obtained running the flowchart of Figure 5-21(a). Numbers represent the order in which each monitor was selected. The colour scale represents entropy (bits). ........................................................................................................ 95

Figure 5-28. Results obtained running the flowchart of Figure 5-21(b). Numbers represent the order in which each monitor was selected. The colour scale represents entropy (bits). ........................................................................................................ 96

Figure 5-29. Entropy values before, at and after the main tributaries ........................................................................................................ 97
Figure 5-30. Solutions obtained by different methods, Total Correlation – Joint Entropy plane. ................................................................. 98
Figure 5-31. Entropy maps for different values of $a$, Eq. (5-1)................................. 99
Figure 6-1. Variation of the Value of Information when changing the prior probability $\pi_p$ and the conditional probabilities $q_{m,s}$ for the consequence matrix shown in Table 6-6............................... 107
Figure 6-2. Definition of $V_x(x)$ and $V_y(y)$........................................................................ 109
Figure 6-3. Definition of $V_x(y)$, $V_z$ and $V_{IOx}$ for a monitor located at $x$ to give the state of the system at $y$ for infinite (a) and finite (b) number of calculation points $y$......... 109
Figure 6-4. Value of two monitors.................................................................................. 110
Figure 6-5. Selection of the best monitors out of three possibilities ......................... 111
Figure 6-6. VOI-related areas to optimise the monitor locations $a$ and $b$, Eq. (6-4)...... 111
Figure 6-7. VOI-related areas to optimise the monitor locations $a$, $b$ and $c$, Eq. (6-5). . 112
Figure 6-8. Definition of flood levels for the canal-pump experiment compared to the minimum, mean and maximum water levels obtained by the model......................... 113
Figure 6-9. Prior beliefs $\pi_p$ estimated with the Table 6-1 ............................................... 114
Figure 6-10. Mean of Eq. (6-2) for all $x$ in the water system and zoomed curve for the no-flood area ........................................................................ 115
Figure 6-11. $V_x$ curve for the calculation point with the highest $V_{IOx}$ (point 43).......... 115
Figure 6-12. $V_x$ curves for the monitors 37 and 57, after solving Eq. (6-4)............... 116
Figure 6-13. $V_x$ curves for the monitors 2, 50 and 73, after solving Eq.(6-5)............. 117
Table 6-8. Definition of the situations for estimation of $q_{m,s}$ for the Magdalena River case............................................................................................................................. 118
Figure 6-14. VOI and the effect of lagged time series..................................................... 119
Figure 6-15. Mean Value of Information estimated with Eq. (6-2) for different values of celerity in the Magdalena River, using a state threshold definition of 80%.......... 120
Figure 6-16. Mean Value of Information estimated with Eq. (6-2) for different values of celerity in the Magdalena River, using a state threshold definition of 50%........... 121
Figure 6-17. Mean Value of Information estimated with Eq. (6-2) for different values of celerity in the Magdalena River, using a state threshold definition of 20%......... 122
Figure 6-18. Results for one, two and three monitor locations for different celerity values and 80% state threshold definition......................................................... 123
Figure 6-19. Results for one, two and three monitor locations for different celerity values and 50% state threshold definition......................................................... 124
Figure 6-20. Results for one, two and three monitor locations for different celerity values and 20% state threshold definition......................................................... 125
Figure 6-21. Mean VOI in the Pijnacker water system, with simplified inputs ............. 127
Figure 6-22. Location of one (a) and two (b) monitors for the Pijnacker water system. 127
Figure 6-23. Selected subsystem of the Pijnacker polder system................................. 128
Figure 6-24. Definition of the possible states, land uses and damage function (consequences)....................................................................................... 129
Table 6-11. Situations .............................................................................................................. 130
Table 6-12. $q_m,s$ matrix for the on-status of the pump downstream $y$ according to the situations of Table 6-11.............................................................................................................. 131
Table 6-13. $q_m,s$ matrix for the off-status of the pump downstream $y$ according to the situations of Table 6-11.............................................................................................................. 131
Figure 6-25. $VOI_t$ maps considering different data sets ..................................................... 132
Figure 6-26. Results for the selected subsystem of the Pijnacker water system using calculated prior beliefs .............................................................................................................. 133
Figure 7-1. Flowchart describing the MoMoX general procedures....................................... 138
Figure 7-2. Technology behind MoMoX................................................................................ 139
Figure 7-3. MoMoX website showing gauge 8 info, and the current water level graph. 139
Figure 7-4. Gauges with scales in cm (a), dm (b) and m (c).................................................... 141
Figure 7-5. Examples of validation errors related to gauge scale......................................... 142
Figure 7-6. Example of error related to datum without negative sign.................................... 143
Figure 7-7. Example of error related to adding an unnecessary negative sign..................... 143
Figure 7-8. Example of random errors due to differences in appreciation......................... 144
Figure 7-9. Water level charts obtained with the data sent by the residents and other participants during the second stage of the experiment .............................................. 145
Figure 7-10. Description of the area, absolute elevations and location of hydraulic structures ......................................................................................................................... 146
Figure 7-11. Zero-model, SMS data and retrieved patterns for the 6 gauged points.............. 148
List of tables

Table 3-1. Characteristics of the pump stations in the Pijnacker polders......................... 33
Table 4-1. Mean hydraulic slope by sectors, Magdalena River........................................ 38
Table 4-2. Number of days of 1995 with discharge and stage data and datum of gauges................................................................. 46
Table 4-3. Characteristics of the grouped wetlands.......................................................... 51
Table 4-4. Resistance number (Manning coefficient) at stations ..................................... 54
Table 5-1. Summary of monitors obtained by each dependency criteria and corresponding values of joint entropy and total correlation........................................... 69
Table 5-2. Number of solutions for approach b) with minimum distance violations by pumps and weirs................................................................. 81
Table 5-3. Sensitivity analysis for parameter u. ............................................................... 85
Table 5-4. Sensitivity analysis for parameter q. ............................................................... 86
Table 6-1. Definition of the vector $\mathbf{\pi}$ for two possible states of a water system ........ 105
Table 6-2. Possible situations of messages at $x$ for given states at $y$ for the case of two possible states................................................................. 105
Table 6-3 Definition of conditional probabilities $q_{m,s}$ according to the situations presented in Table 6-2................................................................. 105
Table 6-4. Definition of the $C_{ac}$ matrix...................................................................... 106
Table 6-5. Definition of actions, states and messages for the canal-pump case.............. 113
Table 6-6. Consequences of doing action $a$ given state $s$ (costs units)......................... 114
Table 6-7. Consequences of doing action $a$ given state $s$ (costs units)......................... 117
Table 6-8. Definition of the situations for estimation of $q_{m,s}$ for the Magdalena River case................................................................. 118
Table 6-9. Definition of conditional probabilities $q_{m,s}$ according to the situations presented in Table 6-8................................................................. 118
Table 6-10. Table of consequences $C_{ac}$ for different land uses for the Pijnacker region 129
Table 6-11. Situations ................................................................................................. 130
Table 6-12. $q_{m,s}$ matrix for the on-status of the pump downstream $y$ according to the situations of Table 6-11................................................................. 131
Table 6-13. $q_{m,s}$ matrix for the off-status of the pump downstream $y$ according to the situations of Table 6-11................................................................. 131
Table 7-1. Description of MoMoX stages .................................................................. 137
Table 7-2. Parameters of the zero-model and SMS model, all dimensionless except V3 to V5 (m$^3$/s).................................................................................. 147
Table 7-3. Parameters of the patterns that fit better the SMS data ............................. 148
Notations

- $a$ Quantization coefficient
- $H(X)$ Entropy of the random variable $X$
- $I(X;Y)$ Mutual information or transinformation between random variables $X$ and $Y$
- $H(X,Y)$ Joint entropy of the random variables $X$ and $Y$
- $C(X,...,Z)$ Total correlation among the random variables $X,...,Z$
- $DIT_{x,y}$ Directional Information Transfer of a variable transmitted from $x$ to $y$
- $T$ Dependency matrix built with a given pairwise criteria
- $v$ Dependency vector given by a row or column of matrix $T$
- $x_q$ Quantized value of $x$
- $u(a,p)$ Utility of the action $a$ chosen with a probability $p$ about the state of the system
- $c_{as}$ Consequences of performing the action $a$ when the system has a state $s$
- $q_{s,m}$ Conditional probability of receiving the message, $m$, given the state, $s$
- $\pi_s$ Decision-maker’s prior probability about the state of the system
- $\pi_{s,m}$ Decision-maker’s posterior beliefs after receiving additional information
- $V_x(y)$ Value of a monitor located at $x$ that provides messages about any other point $y$
- $V_x$ Value of Information curve of a monitor located at $x$ about the entire water system
- $VOI_x$ Value of Information of a monitor located at $x$ about the entire water system
### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIT</td>
<td>Directional Information Transfer</td>
</tr>
<tr>
<td>EPM</td>
<td>Existing Pump Monitors</td>
</tr>
<tr>
<td>HD</td>
<td>Hydrodynamics</td>
</tr>
<tr>
<td>HS</td>
<td>Hydrosheds (hydrologically-corrected version of SRTM elevation data)</td>
</tr>
<tr>
<td>ICT</td>
<td>Information and Communication Technologies</td>
</tr>
<tr>
<td>IDEAM</td>
<td>Institute for Hydrology, Meteorology and Environmental Studies, Col.</td>
</tr>
<tr>
<td>IT</td>
<td>Information Theory</td>
</tr>
<tr>
<td>JH</td>
<td>Joint Entropy</td>
</tr>
<tr>
<td>masl</td>
<td>Meters above sea level</td>
</tr>
<tr>
<td>MOGA</td>
<td>Multi Objective Optimisation with Genetic Algorithms</td>
</tr>
<tr>
<td>MoMoX</td>
<td>Mobile Monitoring Experiment</td>
</tr>
<tr>
<td>MOOP</td>
<td>Multi Objective Optimisation Problem, see section 5.3.2</td>
</tr>
<tr>
<td>NSGA</td>
<td>Non-sorted genetic algorithm</td>
</tr>
<tr>
<td>RR</td>
<td>Rainfall runoff</td>
</tr>
<tr>
<td>SMS</td>
<td>Short Message Service</td>
</tr>
<tr>
<td>SRTM</td>
<td>Shouttle Radar Topographic Mission</td>
</tr>
<tr>
<td>VOI</td>
<td>Value of Information</td>
</tr>
<tr>
<td>WMO</td>
<td>World Meteorological Organisation</td>
</tr>
<tr>
<td>WMP</td>
<td>Water Level Monitoring Design in Polders, see section 5.2.1</td>
</tr>
</tbody>
</table>
Acknowledgements

From the scientific side of this research, I recognise the support of two persons during these years. In the first place, Prof. Roland K. Price, for whom I have two main reasons to thank apart from his continuous guidance and support: his capacity to keep my motivation at a high level during the last few years and the confidence he built in me to successfully perform independent research. I shall always be grateful with him, especially because his teaching has made me a better person. In the second place, the supervision of Dr. Arnold Lobrecht was key because he always facilitated my research path in many aspects, including financial, moral and scientific. I would also like to express my gratitude to the thesis committee members for their interest and valuable comments on my work. On the financial side, I am very grateful to the Delft Cluster and Hoogheemraadschap van Delfland for providing the funds necessary for this research.

The methods and theories developed in this thesis have been applied to two case studies, the polders of Pijnacker and the Magdalena River (Colombia). This would not had been possible without the support of many people. In first place, I need to recognise and thank the Delfland Waterboard staff for their support, in particular Ir. Jan Dragt as the person responsible for opening and maintaining the connections required with the Waterboard and for facilitating the interesting discussions we had during the research; Ir. Job van Dansik for his valuable inputs and positive criticism during my presentations; Ir. Peter Hollanders for facilitating all the information I needed and for stimulating discussions. Many thanks also to Ir. Laura Haitel, who guided me through the flat, beautiful landscape of the polders of Pijnacker and helped me to understand how this complex system works; to Frank Keijzer, who performed the labelling of the gauges used by the MoMoX experiment and provided additional useful information, and to Arie Boele, who provided me with maps of the gauges and other information about the water system.

I feel blessed for the enthusiasm and assistance I received during my time collecting data in Colombia. I specially thank Ing. Paulino Galindo, from Cormagdalena, who always supported this project by providing all the information I asked for. I am especially grateful to Ing. Eduardo Bravo, who introduced me to the secrets of the Magdalena River at the very early stages of my professional career. His experience, stories and ideas about how the River should be managed always inspired me; and this thesis was the perfect excuse for me to get to know more about our River.

The people from the Laboratory of Hydraulics (LEH) of the National University of Colombia, one of the institutions in charge of performing studies in the Magdalena River, have been of great help to me. In particular, I would like to thank Ing. Rafael Ortiz, Ing. William Perdomo, Ing. Cristián Plazas, Ing. Marcela Rodríguez, and Ing. Pedro León, who provided the vital links between Colombia and Delft. My visit to Barranquilla for data collection was so short for collecting the amount of data I needed, that it would not be possible through the kind support of Ing. Manuel Alvarado, director of the Laboratory of Hydraulics of Las Flores (LEH-LF); I am also grateful to Ing. Holbert Corredor, a very experienced engineer who knows the problems and the river shape in its lower zone very
well and with whom I had the opportunity to have a very valuable discussion in Barranquilla. I want to thank Mrs. Myriam Mercado, who was in charge of collecting and storing the data I needed. During my visit to Barrancabermeja, I had the kind support of Ing. Martha Isabel Gualdrón and Ing. Claudia Patricia Guevara, who provided me with many of the documents referenced in Chapter 4. Finally, I must acknowledge Ing. Alvaro Sanjinés in Bogotá and specially Ing. Jorge Enrique Saenz, who shared with me his valuable experience of the Magdalena River.

I have had the fortune to guide two MSc students that supported this research thesis: The first was Liyan He, who succeeded in the difficult task of building the model for the Magdalena River and applying some of the Information Theory approaches developed in Chapter 5. I recognise Liyan’s ability to get through huge amounts of data in Spanish; The second student was Lasantha Rupasinghe, who worked on applying Information Theory approaches to placing sensors for pollution in water distribution networks. Although his efforts are not explicitly presented in this thesis due to the different nature of the system under consideration, he certainly contributed to the modifications of the developed methods and demonstrated their applicability to different water systems. I thank Liyan and Lasantha for sharing with me their time and efforts.

Regarding the MoMoX experiment, I express my sincere gratitude to all the people who supported me in different ways, among others, the MSc students of Hydroinformatics 2009-2011, who participated in the pilot tests of the experiment; Ewoud Kok, who assisted me on advertising the experiment in UNESCO-IHE’s screens and the ladies from the Reception Desk, who took care of the MoMoX-advertisment umbrella for some weeks. The experiment of 22 May was successful thanks to the participants that identified themselves with the following nicknames: nathasja, vladyma, CarlosV, Arlex, angela, Juliette, Gaetano, Aleyda, Leo, Erica_Nino, Marwa, Mesgana, Carolina and odsurein. In the same way, for the experiment held in July in Pijnacker, I thank the residents identified by the nicknames josa, 2641, Rein, Eric, Joanne and Roos. In particular, I thank Rein and Eric, who actively participated on the experiment and gave invaluable feedback through telephone interviews.

I am thankful to two Dutch persons who were particularly important for the completion of this dissertation. First, Ir. Steven Weijs, with whom I held discussions from the very early stages of this research, especially regarding Information Theory. Second, Rosemarijn te Horst, who kindly supported the MoMoX experiment in regard to communicating with people living in the polder, visits to the Pijnacker region, and translations of handouts in Dutch.

I am also grateful to Prof. Dimitri Solomatine, Dr. Andreja Jonoski, Dr. Ioana Popescu, Dr. Biswa Bhattacharya and the entire Hydroinformatics core at UNESCO-IHE. I particularly thank Ir. Jan Luijendijk, Ir. Judith Kasperma, Dr. Schalk Jan van Andel, Dr. Arnold Lobrecht as well as Dr. Peter Kelderman, from Environmental Resources Department, for helping me with the Dutch translation of the summary and propositions. I enjoyed the day-by-day discussions with all of them.
Similarly, I want to express my gratitude to my PhD colleagues, including of course the glorious Tax-Free Employees football team (we are the champions) and those with which I spent time discussing some of the topics of this thesis: Solomon Seyoum, Michael Siek, Durga LaL Shrestra, Fikri Abdulah, Ivan García, Carolina Rogeliz, Mónica Sáenz. My appreciation goes to all my IHE colleagues that I have known during these years at UNESCO-IHE.

I need to thank also my colleagues and friends Carlos Velez, Arlex Sánchez, Wilmer Barreto and Gerald Corzo for their feedback on my research during our lunch breaks and their inputs to motivate my creativity, not only for my PhD but for my entrepreneur ambitions. I am grateful to my friends Fernando, Claudia, Zaira, Javier, Roberto and Berenice, who were very close to me at the early stages of my PhD and have supported me during this process; also to Angela, Gaetano, Carmen, Carlos, Diego, Carol, Gianluca, Diana, Natashja and Ruzika, who I consider as the extension of my family in The Netherlands. Especial thanks go to Patricia Nieto, my mother in law, who indirectly helped me to finalise this document.

Last but not least, I would like to thank my family, father, mother and brothers for their support from the distance. Above all, my gratitude for my beloved wife, Sandra, who has always supported me unconditionally during all the stages of this research. Cosita, gracias por tu apoyo y paciencia durante estos últimos años. Este es también tu logro. Agradezco tu sacrificio y te agradeceré siempre por esos dos seres tan maravillosos que son Valentina y Sebastian.
About the author

José Leonardo Alfonso Segura was born in 1974 in Bogotá, Colombia. In 1999 he graduated as Civil Engineer from the Faculty of Engineering of the National University of Colombia, where his father taught Fluids Mechanics and Hydraulics for 35 years. After his graduation, and influenced by his father’s steps, he entered the LEH Laboratory as a Junior Engineer, in the same University. There he worked on projects concerning the Magdalena River that included navigation path designs, dredging plans, dynamics of river morphology, flood and erosion control assessments. At LEH Mr. Alfonso developed his first PC routines to speed up certain tasks, in particular for data processing.

By the end of 2001 he decided to move to the private sector, working for Aquadatos, a consultancy company that specialises in mathematical modelling of water distribution networks and urban drainage systems. There he used his programming skills to create diverse tools to facilitate data processing, editing drawings and optimising software usage. In 2003 Mr. Alfonso was promoted to Technical Director, and he continued to develop ICT tools for the company, while managing groups of junior engineers in diverse projects, the most important of which was the development a methodology for assessing the supply networks of 20 capital cities of Colombia using mathematical models. His strong affinity for water and computers led him seek for further qualifications and to look for specialization courses in which both aspects were combined.

He found at UNESCO-IHE, Delft exactly what he was looking for, and in 2004 he started his MSc studies in Hydroinformatics with a fellowship granted by the Dutch Government’s Watermil Project, and graduated in 2006 with a thesis that explored the use of different Hydroinformatics tools for real time management of water quality in distribution networks, with a case study in Villavicencio, Colombia. Although not in his original plan when he first arrived in Holland, he decided to continue with PhD studies, now working on the challenging task of developing approaches for designing and evaluating monitoring networks. His findings are presented in this thesis.

List of publications


Samenvatting

Het meten van de verschillende processen van de hydrologische kringloop is uitermate belangrijk voor een efficiënt waterbeheer. Meetnetten bieden beheerders gegevens voor de analyse van het historische en het actuele waterbeheer, waarmee zij informatie beschikbaar krijgen om beslissingen te nemen over hun watersysteem. Belangrijke uitdagingen bij het ontwerpen en evalueren van meetnetten zijn o.a. het vinden van de juiste tijd- en ruimteschalen en de gewenste dekking tegen de laagste kosten.

Het theoretisch optimale meetnetwerk is al langere tijd onderwerp van wetenschappelijk onderzoek; toch blijken de meetgegevens die met de bestaande meetnetwerken worden verkregen onvoldoende te zijn om de dynamica van natuurlijke systemen te verklaren. Dit komt wellicht omdat de criteria voor het vaststellen van het meetnetwerk veelal worden bepaald door niet-wetenschappelijke, maar door politieke en sociale factoren. Nieuwe benaderingen, waarbij de aard van de te nemen beslissing en de stem van de belanghebbenden worden meegenomen, kunnen een brug slaan tussen theorie en praktijk.

In dit promotieonderzoek, dat door Delft Cluster, het Hoogheemraadschap van Delfland en UNESCO-IHE is gefinancierd, worden innovatieve methoden voor het ontwerpen en evalueren van meetnetten behandeld. De leidende gedachte hierbij is het ontwikkelen van een optimaal functionerend watersysteem op basis van optimale informatie uit het meetnet in dat watersysteem. Als we spreken over het functioneren van een watersysteem, bedoelen we de classificatie van het gedrag van het water met betrekking tot iedere specifieke vorm van watergebruik en belang binnen een watersysteem. De mate van functioneren van het watersysteem zou bepalend moeten zijn voor de beslissingen die over het watersysteem worden genomen. Als we spreken over het maximaliseren van de informatie, dan doelen we op het vinden van die mogelijke locaties waar metingen de beste indicatie geven van de toestand van ieder punt in het watersysteem.

Dit onderzoek is gebaseerd op drie pijlers: Informatietheorie, de Waarde van Informatie (Value Of Information; VOI), en gegevensverzameling door publieke participatie. De eerste pijler is een methode waarmee de locaties van waterstandmetingen worden vastgesteld op grond van de informatietheorie, met nadruk op het reduceren van onzekerheid. Met behulp van de Informatietheorie van Shannon is het gelukt om de meetlocaties voor waterstanden in een watersysteem zodanig te bepalen dat de onzekerheid wordt verminderd. Er zijn twee methoden ontwikkeld. Als eerste de Water Level Monitoring Design in Polders (WMP) methode, waarin de meetinstrumenten één voor één zodanig worden geplaatst dat de inhoud van informatie van de set waterstandmetingen zo groot mogelijk blijft, terwijl de meetinformatie die door de meetinstrumenten wordt gedeeld zo klein mogelijk blijft. De tweede methode gaat uit van een Multi-Objective Optimization Problem (MOOP), waarin twee praktische overwegingen worden afgewogen: 1) de kosten van het plaatsen van nieuwe meetapparatuur, en 2) de kosten van het te dicht bij een kunstwerk plaatsen van meetinstrumenten. De kosten zijn uitgedrukt in informatie-enheden, waarvoor extra termen in de doelfunctie van het optimalisatieprobleem werden geformuleerd. Voorts is
de MOOP methode uitgebreid met een ‘rank-based greedy’ algoritme en toegepast voor het bepalen van de locatie van afvoermetingen in de Magdalena rivier.

De tweede pijler behandelt een methode om meetinstrumenten te plaatsen op grond van de waarde die de gebruiker toekent aan de ingewonnen informatie voor het nemen van beslissingen. Hiervoor zijn concepten van de theorie van de ‘Waarde van Informatie’ gebruikt. Deze zijn gedefinieerd als het verwachte verschil tussen de opbrengst van het kiezen van een bepaalde actie op basis van de bestaande schatting van de toestand van het watersysteem, en de opbrengst wanneer de actie wordt gekozen met inbegrip van de extra informatie van het meetinstrument. De methode selecteert de meest waardevolle set van meetlocaties voor een bepaald watersysteem op grond van de door de gebruiker vooraf verwachte toestand van het systeem, de gevolgen voor het functioneren van het systeem, en de kwaliteit van de informatie die het meetnet beschikbaar zou kunnen stellen. Deze nieuwe benadering is uitgewerkt door een theorie te ontwikkelen voor het bepalen van de waarde voor het plaatsen van één meetstation en dit uit te breiden voor de bepaling van \( n \) meetstations. Daarnaast wordt voorgesteld de benodigde probabilistische variabelen voor het uitrekenen van de informatiewaarde (VOI) te schatten met behulp van computersimulatiemodellen.

De derde pijler van het onderzoek, met grote praktische waarde, heeft als doel nieuwe mogelijkheden te verkennen voor het verzamelen van informatie met behulp van mobiele telefoons om vervolgens hiermee simulatiemodellen te verbeteren. Vandaag de dag kan een mobiele telefoon niet meer louter gezien worden als een toestel voor het voeren van gesprekken; ze combineren functies zoals PC software, digitale camera, rekenmachine en agenda en leveren Internet, radio, televisie en faxdiensten. Zo kunnen mobiele telefoons ook gebruikt worden voor het waarnemen van waterstanden met hulp van het publiek. Het idee is om gebruik te maken van de voordelen van het betrekken van het publiek bij het doen van waarnemingen. Deze voordelen worden door meerdere auteurs beschreven als: het creëren van maatschappelijke bewustwording van milieuproblemen, het verbeteren van samenwerking tussen belanghebbenden, de mogelijkheid tot kosten-baten analyse van dataverzameling, en het grote bereik in ruimte en tijd. Een belangrijke bijdrage aan dit onderzoek werd geleverd door het ‘Mobile Monitoring Experiment’ (MoMoX), een experiment dat is uitgevoerd in 2010 in de polders van Pijnacker, met deelnemers met verschillende betrokkenheid, waaronder bewoners dichtbij de locaties waar waterstanden worden gemeten.

Twee zeer verschillende case studies zijn gebruikt om de ontwikkelde methoden en theorieën te testen. Voor de locatie van de eerste case study is gekozen voor de polders van Pijnacker, een typisch laaggelegen vlak stuk Nederland met een in hoge mate gecontroleerd watersysteem. Dit gebied is gelegen aan de oostkant van het beheersgebied van het Hoogheemraadschap van Delfland. Het is een plattelandsgebied, met hier en daar wat stedelijke ontwikkeling en broeikassen. Hydrologisch gezien bestaat het gebied uit vier hoofdpolders, onderverdeeld in 127 kleinere onafhankelijke afvoergebieden, elk met een eigen doelstelling voor het peilbeheer. De waterstanden in het systeem van kanalen worden tussen bepaalde niveaus geregeld door een stelsel van gemalen, inlaten en stuwen.
De tweede case study betreft de Magdalena rivier die het belangrijkste riviersysteem van Colombia vormt en de grootste rivier is die uitmondt in de Caribische Zee. De rivier heeft een lengte van 1530 kilometer van zuid naar noord, en voert water af uit een stroomgebied dat 24% van het hele land beslaat en waar 77% van de bevolking woont. Het gebied dat in dit onderzoek is bestudeerd, beslaat het middelste en benedenstroomse deel van de rivier. Dit gebied is niet alleen belangrijk vanwege de economische activiteiten zoals scheepvaart, visvangst en landbouw, maar ook omdat dit gebied het meest getroffen wordt door overstromingen.

De resultaten van dit onderzoek laten zien dat meetnetwerken geëvalueerd en ontworpen zouden kunnen worden op basis van nieuwe criteria, zoals de inhoud van informatie, het soort gebruiker van de informatie en de mogelijkheden die de huidige mobiele technologie biedt voor het verzamelen van gegevens. Er zijn nieuwe methoden ontwikkeld die kunnen worden gebruikt voor het optimaliseren van meetnetwerken, waarbij simulatiemodellen worden gecombineerd met concepten uit de informatietheorie en onderdelen van theorieën rondom de Waarde van Informatie. Tenslotte is een openbaar meetnetwerk voor het verzamelen van waterstanden met behulp van mobiele telefoons ontworpen, getest en beoordeeld.

Naast deze resultaten draagt dit onderzoek nieuwe mogelijkheden aan voor de toepassing van informatietheorie in het waterbeheer, waar informatie normaliter gekwantifieerd wordt in entropie-eenheden. Door gebruik te maken van concepten van de Waarde van Informatie, kunnen zowel de op entropie gebaseerde methoden die zijn behandeld in dit onderzoek, als de methoden die zijn ontwikkeld in eerder onderzoek, aangepast worden, zodat ze ook meetinformatie aan geld kunnen relateren.
Monitoring networks provide data that is analysed to help managers make informed decisions about their water systems. Their design and evaluation have a number of challenges that must be resolved, among others, the restriction on having a limited number of monitoring devices.

This book presents innovative methods to design and evaluate monitoring networks. The main idea is to maximise the performance of water systems by optimising the information content that can be obtained from monitoring networks. This is done through the combination of models and two theoretical concepts: Information Theory, initially developed in the field of communications, and Value of Information, initially developed in the field of economics. Additionally, the possibility of using public participation to gather information with mobile phones to improve models is also explored in the research. Two very different case studies are used to test the developed methods and theories: Pijnacker, a typical low-lying regional polder in the Netherlands, which is highly controlled and the Magdalena River, the major river system in Colombia.

The results of this research demonstrate that monitoring networks can be evaluated and designed by considering new variables, such as the information content, the user of the information and the potential of current mobile phones for data collection.