Data assimilation in hydrological modelling

Drecourt, Jean-Philippe

Publication date:
2004

Document Version
Publisher's PDF, also known as Version of record

Environment & Resources
Technical University of Denmark

Data assimilation in hydrological modelling

Jean-Philippe Drécourt

Data assimilation in hydrological modelling

Jean-Philippe Drécourt

June 15, 2004

Data assimilation in hydrological modelling

Cover: Birte Brejl
Printed by: DTU tryk
Environmental & Resources DTU
ISBN 87-89220-84-6

The thesis will be available as a downloadable pdf-file from the department's homepage on: www.er.dtu.dk

Environment & Resources DTU
Library
Bygningstorvet, Building 115, Technical University of Denmark
DK-2800 Kgs. Lyngby

Phone:
Direct (+45) 45 25 16 10
(+45) 45 25 16 00
Fax: (+45) 45 93 28 50
E-mail: library@er.dtu.dk

L'exact pris pour le vrai.
(Exactness mistaken for the truth.)
Victor Hugo

Abstract

Data assimilation is an invaluable tool in hydrological modelling as it allows to efficiently combine scarce data with a numerical model to obtain improved model predictions. In addition, data assimilation also provides an uncertainty analysis of the predictions made by the hydrological model. In this thesis, the Kalman filter is used for data assimilation with a focus on groundwater modelling. However the developed techniques are general and can be applied also in other modelling domains.

Modelling involves conceptualization of the processes of Nature. Data assimilation provides a way to deal with the uncertainties resulting from the model creation and calibration. It is necessary to balance modelling uncertainties and observation uncertainties to prevent an excessive forcing of the model towards the observations.

The popularity of the Kalman filter resulted in the development of various techniques to deal with model non-linearities and biased errors. A literature review analyzes the most popular techniques and their application in hydrological modelling.

Since bias is an important problem in groundwater modelling, two bias aware Kalman filters have been implemented and compared using an artificial test case. It resulted in the recommandation of the Colored Noise Kalman filter as the most suitable method. By using bias feedback in the model propagation, the bias variations are represented by their first order approximation.

The main contribution of this thesis is the development of a sequential calibration technique whereby the performance of the model and its associated Kalman filter is optimized separately from the uncertainty analysis. Instead of using rules of thumb to estimate the parameters of the covariance matrices, the method relies on an objective automatic calibration method that aims at optimizing the performance without affecting the stability of the system. The application of the technique to an artificial case leads to a Kalman filter setup that generates a minimum overall model error as well as an optimized uncertainty analysis.

The sequential calibration scheme has been further developed for the simultaneous calibration of the Kalman filter and the physical model parameters. The procedure was applied to the Danish Karup catchment. It resulted in a significant reduction of the model error and an optimized uncertainty estimation both at assimilation and validation points. However, the analysis showed that care should be taken in the calibration since some parameters may lack physical interpretability.

Resumé

Data assimilering er et uvurderligt redskab i hydrologisk modellering idet det muliggør en effektiv kombination af spredte data med en numerisk model til at opnå forbedrede model prediktioner. Derudover giver data assimilering en analyse af usikkerheden på prediktionerne af den hydrologiske model. I nærværende afhandling er Kalman filteret benyttet til data assimilering med fokus på anvendelse i grundvandsmodellering. De udviklede teknikker er dog generelle og kan benyttes også i andre model domæner.

Modellering involverer konceptualisering af de processor der optræder i naturen. Data assimilering giver mulighed for at håndtere usikkerheder i forbindelse med model opbygning og kalibrering. Det er nødvendigt at balancere model usikkerheder og usikkerhederne på observationerne for at undgå at modellen i for høj grad trækkes mod observationerne.

Kalman filterets popularitet har resulteret i udvikling af forskellige teknikker til håndtering af model ikke-lineariteter og fejl der har en bias. En gennemgang af litteraturen analyserer de mest populære teknikker og deres anvendelse i hydrologisk modellering.

Idet bias er en vigtig problematik i grundvandsmodellering er to forskellige Kalman filter teknikker med bias korrektion blevet implementeret og sammenlignet i et syntetisk model setup. Det resulterede i en anbefaling af Kalman filter med farvet støj som den mest anvendelige metode. Ved benyttelse af feedback af bias i model propageringen bliver bias variationerne repræsenteret ved en første ordens approksimation.

Hovedbidraget i denne afhandling er udviklingen af en sekventiel kalibreringsmetode hvor Kalman filteret først er optimeret med hensyn til modellens prediktionsevne og dernæst med hensyn til usikkerhedsanalysen. I stedet for brug af tommelfingerregler til estimering af parametrene i kovarians matricerne beror metoden på en objektiv automatisk kalibreringsmetode som har til formål at optimere modellens prediktionsevne uden at det har en negativ effekt på stabiliteten af systemet. Anvendelsen af metoden i et syntetisk test giver et Kalman filter setup som set over hele modeldomænet har en minimum model prediktionsfejl og desuden giver en optimeret usikkerhedsanalyse.

Den sekventielle kalibreringsrutine er blevet yderligere udviklet for samtidig kalibrering af Kalman filteret og parametrene i den fysiske model. Denne procedure er blevet anvendt på en opsætning af det danske Karup opland. Det resulterede i en betydelig reduktion af model prediktionsfejlen og en optimeret estimation af prediktionsusikerheden i både assimileringspunkter og valideringspunkter. Analysen viste dog at man skal vre varsom med kalibreringen idet man kan få parametre som ikke er fysisk realistiske.

Foreword

The present thesis is prepared as one of the requirements for the Ph. D. degree. The thesis is composed of a general discussion about data assimilation with a focus on hydrological modelling and four adjoined papers that represent the main milestones of the study.

The major part of the study was carried out at the River and Flood Management department in DHI Water & Environment, Hørsholm, Denmark under the supervision of Henrik Madsen and at the Environment & Resources department of the Technical University of Denmark, Lyngby, Denmark under the supervision of Dan Rosbjerg. Their support and guidance are gratefully acknowledged. Two months were spent at Delft Technical University, The Netherlands, under the supervision of Arnold Heemink, whose support is acknowledged.

The study was in part funded by the Danish Technical Research Council (STVF) under the Talent Project No. 9901671 entitled "Data Assimilation in Hydrological and Hydrodynamic Modelling"

Hørsholm, February 2004

Jean-Philippe Drécourt

The papers are not included in this www-version but can be obtained from the Library at Environment & Resources DTU, Bygningstorvet, Building 115, Technical University of Denmark, DK-2800 Lyngby (library@er.dtu.dk).

Contents

Chapter 1

Introduction

1.1 Background

The discovery of scientific knowledge, either in the form of rules, empirical experience or mathematical equations always relies on data. Even in mathematics, a science normally considered as abstract and relying only on deduction, the basic axioms are linked to observations of the world. These observations are quantified, measured and then conceptualized so that they reach a level of abstraction that can be manipulated by the reasoning tools that constitute mathematics. The resulting equations, laws or theorems apply in an idealized world but provide an insight to the real world in which we live, and provide tools to predict its behavior.

The use of models that are defined by a set of mathematical equations is the preferred way for physical scientists to predict the future of systems of all sizes, from a few subatomic particles to the working of the entire universe, and even beyond. Hydrology is not an exception. From the prayers of the priests in the ancient Egypt to predict the Nile floods to the modern computerized flood forecasting and water resources, the behavior of water has been observed, measured, conceptualized and then modeled.

Until the beginning of the XX^{th} century, it was believed that models could become so precise and so complex that it would be possible to predict the behavior of any system, simply by calculating for long enough. The use of mechanic calculation and the improvement of scientific knowledge made people extremely confident in the power of mathematics and what we call today *deterministic modelling* to reach any level of precision. The impact of the discovery of quantum mechanics on the way the world was perceived, reached beyond the restricted circle of nuclear physics. The change in mentality reached all the modelling scales. It became clear that we would not be

able to describe the world at any level of precision. The simple fact of observing would affect the system because instrumentation was not "invisible" to the system, and the models that we were so proud of proved to be too simple to represent the behavior of a world that is in all its aspects *fractal* and *chaotic*.

This fundamental dichotomy between the different models of the world and the world itself is strengthened in hydrology by another problem: we do not and we cannot know what exactly happens in hydrological systems: in most cases, it is impossible to get a clear picture of the whole system either because it is hidden like in groundwater, or the system is so complex that we cannot model it without making assumptions that have negative impact on the accuracy of the model. This results in sparse data and overly simple models. The attempts to solve this problem take two opposite paths.

With the increasing power of computers, the models are made more complex by increasing the number of physical processes that are represented. Since observations do not provide sufficient information about the detailed processes that are represented, it is impossible to calibrate the model properly.

The opposite approach is encompassed in the term *data mining*. The original philosophy behind data mining is the attempt to circumvent the physical models. If the model is too rough to represent reality, or impossible to calibrate, why not rely entirely on the data and use some mathematical interpolation and/or extrapolation where data are not available. In domains where data have been collected over long periods, the scientists using data mining hope to capture and reproduce the true dynamics of the system just by analyzing the data. Some data mining techniques also incorporate physical knowledge in the data oriented model building. These approaches are very successful in domains where the physical models are poor or do not exist but they discard the experience accumulated by hundreds of years of refinement of theories and ignore to a certain degree the risk of overfitting a model to data that are corrupted by different types of observation error.

1.2 The discovery of data assimilation

Data assimilation started as a military project to control the trajectory of missiles in the 60s. Using a model alone would lead to erroneous trajectories because of the incomplete knowledge of the atmospheric conditions and it was impossible to get data accurate enough to rely solely on them. Therefore a method was designed to to take the best of both worlds: where there is no observation, a physical model is used and relied upon. Where good

data are available, they are used to represent the system, and above all, the uncertainty on both the data and the model are taken into account.

Data assimilation has been used in different domains and recently became successfully used in earth sciences like atmospheric modelling, oceanography and hydrology. The upcoming of remote sensing data and the ever-increasing power of computers made possible the use of large but simple models where data from satellites proved a valuable addition to the relative simplicity of the model.

1.3 Outline of the thesis

This thesis is divided into three parts. The first part consists of a general discussion about the assumptions made with modelling. It explains how data assimilation is the proper tool to take into account the imperfection of the model during calibration. It focuses on hydrological modelling and in particular on groundwater modelling. This part includes also a presentation of the papers provided in the appendix and a brief discussion of the motivations to explore the methods presented here.

The second part is a presentation of the Matlab toolbox that has been developed during the study to ease the manipulation of time series, ensemble simulation and state/space representation.

Finally, the papers that summarize the essential steps of the research are compiled. There are four papers:

- **Paper A**: Kalman filter in hydrological modelling, by J. P. Drécourt, technical report available at `http://projects.dhi.dk/daihm/Files/KFlitreview.pdf`.
- **Paper B**: Bias aware Kalman filters:Comparison and improvements by J. P. Drécourt, H. Madsen and D. Rosbjerg, submitted to *Advances in Water Resources*.
- **Paper C**: A calibration framework for a Kalman filter applied to a groundwater model by J. P. Drécourt, H. Madsen and D. Rosbjerg, submitted to *Advances in Water Resources*.
- **Paper D**: Joint calibration and uncertainty analysis of a groundwater model coupled with a Kalman filter by J. P. Drécourt, H. Madsen and D. Rosbjerg.

Chapter 2

General discussion

2.1 From Nature to the Model

Hydrological systems and in particular groundwater systems are complex and difficult to observe. The amount of observations available is limited in time but especially in space for two reasons:

- As wells have to be dug to observe the piezometric heads, their location is restricted to where it is possible to set them up. The information obtained through observation is therefore punctual;
- The observation process itself is disruptive: not only the presence of the well generates a small decrease in the piezometric head but it is also disruptive to the structure of the soil and can lead to water movements that did not exist before the digging of the well.

Therefore the model is an indispensable part of the monitoring of groundwater resources. It acts as an interpolator at locations where it is practically impossible to observe the necessary information. This section will give a brief overview of the processes of translation of real world processes into equations that can be used in a computer.

2.1.1 Collection of data

This section reviews in general terms the data needed to get a good picture of a groundwater model system.

Geological data

Groundwater systems are extremely complex water systems, because compared to other hydrological systems, the main source of influence is not the

water itself but the soil and rocks it evolves in. Therefore to generate a good model, there is not only the need to know how the water behaves but also the geological property of the aquifer. As with the piezometric heads, the information is hidden and difficult to access. The main information needed are the physical properties of the soil, as they will influence the flow of the water. The geological structure of the aquifer provides information about the hydraulic properties of the soil. The issue with geological information is its small scale variability and heterogeneity that can influence greatly the flow of water by creating preferential paths where the water can flow much faster and affect the way piezometric heads evolve through time. In some geological formations, fractures are also responsible for the majority of the water flow and will affect the predictions of a groundwater model. These fractures are extremely difficult to detect during geological surveys and are one of the great unknowns of groundwater modelling.

Forcing data

Geological data provide information about the way a groundwater system reacts to forcing. The forcing is mainly constituted of recharge. The recharge can come from different sources:

- The rainfall that is usually the main source of recharge
- Groundwater inflow from the boundary of the catchment or the system study
- River inflow
- Negative recharge in the form of for example drainage, pumping or evapotranspiration.

Each of the components listed above are also difficult to measure. One of the main studied issues is rainfall [18, 15, 22]. The error on measurement of rainfall can be divided into three different categories:

- An underestimation due to wetting of the walls and evaporation in the rain gauge;
- An underestimation due to wind dispersion of the rainfall around the rain gauge;
- A measurement error;

Rainfall is also a spatially heterogeneous process: the use of point-measurements to extrapolate the rainfall over a whole catchment increases the uncertainty that is already large.

The other sources of recharge are even more difficult to measure. For example evapotranspiration is normally estimated from physical models that use leaf area and temperature that are known to be rather inaccurate.

2.1.2 Conceptualization of reality

When developing a model, it is necessary to conceptualize nature, i.e. to simplify the processes that occur in reality. The three main reasons for doing so are:

- The lack of knowledge about certain processes and their interaction
- The lack of mathematical tools available to solve analytically the set of equation and/or differential equations that describe the way the system behaves
- The finite computing power available limiting the size of the data that can be handled

Discretization

The approach that is adopted in hydrological models to allow solving differential equations is to use discretization, that is to turn the differential equation from a continuous representation in time and space into a discrete representation. For time representation, finite difference is generally used, whereas spatial discretization can be done by finite difference or finite elements. Instead of manipulating continuous variables, the model represents space and time by elementary units of volume or time where the variables and parameters are assumed to keep the same value.

Linearization

Discretization can be seen as a particular case of linearization. In the case of discretization, the state can be considered as a piece-wise linear approximation of the behavior represented by the original differential equations. With linearization, the model evolution can actually be represented by its Jacobian. The motivation for linearization is the mathematical simplicity of solving the problem by using its quadratic properties. The most well-known algorithm used is the Preconditioned-conjugate gradient (PCG) [19].

Markov chain

The assumption behind the Markov chain approach is that the state of the system at a given time step depends only on the state of the system at the previous time step, and the forcing.

Associated with the assumption of linearity, and the existence of some external forcing as it is usually the case in hydrology, the model can be written as:

$$\mathbf{x}_{k+1} = \mathbf{M}_k \mathbf{x}_k + \mathbf{u}_k \tag{2.1}$$

where $\mathbf{x}_k$ is the state of the system at time step k, $\mathbf{M}_k$ is the linearized model operator and $\mathbf{u}_k$ is the forcing to the system.

This model form has the advantage of being computationally efficient as it only demands the storage of the previous time step in memory. Most computer models are designed using the Markov chain representation and unless stated otherwise, it is the assumption taken in this document. The linear Markov chain is usually used to design data assimilation methods, because it leads to optimal solutions. More complex models including for example non-linearities lead to suboptimal solutions derived from the linear Markov chain case.

Conceptual model

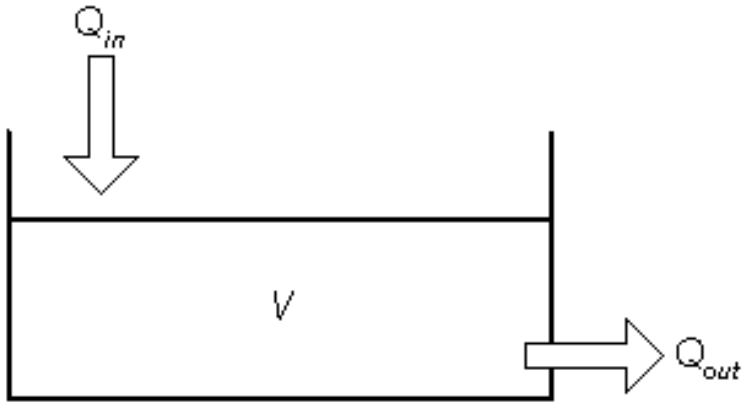

Figure 2.1: Schematic representation of a linear reservoir. Q_{in} is the flow into the reservoir [L^3T^{-1}], Q_{out} is the flow out of the reservoir [L^3T^{-1}] and V is the volume of the reservoir [L^3]. The reservoir behavior is given by equation 2.2

Any model is at a certain level a conceptual model, as it is always a simplified representation of reality. A model is called conceptual when it describes the system studied using simple mathematical formulations. In hydrology, the most popular conceptual model is the linear reservoir. The linear reservoir behavior (cf. figure 2.1) is governed by the assumption that the flow rate at the outlet of the reservoir is directly proportional to the content of the reservoir [5]:

$$Q_{out} = \frac{1}{C} \cdot V \tag{2.2}$$

where C [T] is the time constant that defines how fast the water flows out of the reservoir. Combinations of linear reservoirs have been used extensively in unsaturated zone modelling to avoid the problems linked to the Richard's equations. For example, the NAM model [11] is a lumped rainfall-runoff model that uses five different linear reservoirs to model how the unsaturated zone and the saturated zone react to rainfall to generate runoff.

Data mining, with for example the use of artificial neural networks, is at the extreme end of conceptualization. The model is a generic extrapolation device that is used to fit the observed data. It is expected that proper training techniques will lead to a model that captures the essential dynamic features of the system studied. In practice, it demands a lot of care, and a large data set to test the model on, to obtain a model that represents the system properly without overfitting.

Simplifying the model: model reduction

Model reduction can be considered as a kind of conceptualization, not of nature, but of a complex physical model. The goal of model reduction is defined in [12] as the:

> Develop[ment of] a general method with which complex computational models can be effectively and efficiently reduced.

The reduction process is divided into five steps:

1. Definition of the information required from the model. This process selects the data and structures that are relevant and should be included in the reduced model.

2. Data processing. The irrelevant data are filtered out.

3. Identification of the most important patterns in the spatial and temporal variations of the model .

4. Formulation of a model that can describe and reconstruct the reduced data sets. The result is the reduced model.

5. Assessment of the reduced model performance for other cases. The reduced model is validated on data that have not been used for the creation of the model.

In a nutshell, the reduced model is a surrogate model. It has generally little to do with the original processes involved in the system and can be considered as a result of data mining technique. The last step of the creation of the reduced model is therefore extremely important as it tests the generalization abilities of the model.

2.1.3 Calibration

For a model to be useful, it should be general enough to be able to fit a large range of problems but also contain some adjustment procedures (generally in the form of parameters) that allow to tailor the model to the specific behaviour of the system studied. For data mining techniques, the category of problems covered by the model can be very broad, while physical models are limited to a more specific range of problems, like unsaturated zone modelling, rainfall-runoff modelling or hydrodynamical modelling.

To make the models more specific to the problem studied, a set of constants (the parameters) are adjusted in order to reach a suitable level of concordance between the model output and the observations of the system. The adjustment is called calibration. The questions raised by the use of calibration are numerous and difficult to answer in a general framework.

- What is the suitable level of concordance between the model and the system observed?
- How to measure the level of concordance?
- Is there one and only one set of optimal parameters?
- What information can be extracted from the resulting parameter set(s)?

Fitting the data

Given the fact that we can measure perfectly the accuracy of a model to reproduce the observed data, how good should the model fit the data? The obvious answer to this question is *perfectly*. But is it actually possible to fit perfectly a model and the data that have been observed? The model is a

simplification of reality, at the best it represents slightly simplified dynamics, relies on data whose accuracy is questionable and that are sparse in time and space. It is therefore unlikely that there exists a set of parameters that can perfectly reproduce the behavior of the system studied.

Since the model is known to be an imperfect representation of reality, a perfect match between the data and the model is not desirable. It would lead to a model that mimics the system without "understanding" it, i.e. the model would not be able to perform one of its major roles: forecasting. The overfitted model reproduces very well the data that have been used for calibration but cannot capture the overall behavior of the system studied.

Overfitting is the plague of modelling and is extremely difficult to detect during the calibration procedure. It should nevertheless be avoided as it turns a model that is capable of representing a system with a certain degree of accuracy into a mirror of the data that provides very little information about the system studied.

The objective function

The simplest approach to calibration, and yet the most difficult to quantify, is the manual calibration using visual judgement of the goodness of fit. This method relies solely on the experience of the person calibrating the model and leads to situations where it is difficult to decide which parameter set is "best". Computers have allowed for the development of automated calibration methods that need more objective ways of defining how good a model is. It is achieved by the use of objective functions.

An objective function is a measure of numerical closeness between the observed data and the modelled data. It necessarily implies an assumption on what sort of closeness is looked for. The root mean square error is probably the most popular fitness measure. It is calculated as:

$$RMSE = \sqrt{\frac{1}{T}\sum_{k=1}^{T}(x_k - y_k)^2} \tag{2.3}$$

where x is the model output, y is the observed data, the index k represents the different time steps at which the system is observed and T the total number of time steps. The closer the RMSE is to zero, the better the model is. However this objective function aggregates information about the shape and the bias of the system. It is indeed possible to decompose the RMSE measure into a bias measure, the average error (AE), and a shape measure, the standard deviation of the residuals (STD):

$$RMSE^2 = AE^2 + STD^2 \tag{2.4}$$

where

$$AE = \frac{1}{T}\sum_{k=1}^{T}(y_k - x_k) \tag{2.5}$$

and

$$STD = \sqrt{\frac{1}{T}\sum_{k=1}^{T}\left((y_k - x_k) - \frac{1}{T}\sum_{k'=1}^{T}(y_{k'} - x_{k'})\right)^2} \tag{2.6}$$

This means that for a given RMSE, the model can be good at correcting bias and badly model the dynamics or vice-versa.

Another example of an objective function is the Pearson r^2 measure. It is calculated as:

$$r^2 = \frac{\sigma_{xy}^2}{\sigma_x^2 \sigma_y^2} \tag{2.7}$$

where σ_{xy}^2 is the squared covariance between the observed data and the output, σ_x^2 is the variance of the model output, and σ_y^2 is the variance of the observations. If $r^2 = 1$, then there is a unique linear relationship between the observations and the model data, i.e., there exist a unique set of constants $a, b \in \mathbb{R}$ with $a \neq 0$ so that for every k:

$$y_k = a \cdot x_k + b \tag{2.8}$$

This measure is generally used as an indicator of how well the model represents the shape of the variation of the observations, though a model that outputs a hydrograph that is identical to observations, given a linear transformation (i.e. $a \neq 1$ and/or $b \neq 0$), is not considered to be a good model.

Calibration is therefore always a subjective process biased by the choice of the objective function.

Equifinality

Equifinality looks at the problem of subjectivity of the calibration in the light of the choice of the parameters. Assuming that the objective function used is an acceptable way of distinguishing between the good and the bad models, in most cases we can obtain parameter sets that are equally good. It has been studied in great details by Beven ([2, 9] for example). His general argument is that over-parameterization of hydrological models and the limited amount

of calibration data available lead to equifinality of possible solutions. Using Monte Carlo simulations, he shows that large range of parameters lead to the same objective function value, i.e. the same performance of the model. The equifinality problem is closely related to the choice of the objective function as two models that are equally good in the RMSE sense can for example be different in the Pearson r^2 sense.

Pareto ranking

It is possible to partially "unfold" the equifinality problem by using Pareto optimization [10]. This technique is used in **Paper C** to get a better overview of the behavior of the system by looking at different objective functions at the same time. The result is a set of non-dominated solutions (cf. **Paper C** for more details about the technique) that have different characteristics and show the tradeoff between the different demands related to each objective function. Even though the information obtained through the calibration process is actually more detailed, the hydrologist is given a set of parameter sets that cannot be distinguished using the objective functions. Here again, the subjectivity of the hydrologist is necessary to choose the supposed best model.

Any information in the parameter set?

The set of parameters resulting from the calibration contains some information and provides some insight about the physical system that is studied. Nevertheless it is important to remember the assumptions made during the modelling approach in order not to extract wrong information about the model.

Calibration of hydraulic conductivities using piezometric head observations is a typical example. When using the usual Darcy flow approach to model an aquifer, the resulting hydraulic conductivity is a macroscopic parameter. A high conductivity does not necessarily mean that the aquifer is highly permeable but that water, in a way or another, has a high velocity. It can be through fractures in a low conductivity media, or in a highly heterogeneous media with preferential flow. Only additional observations like tracer concentration can lead to better interpretation of the system.

Special caution is needed with conceptual models. The NAM rainfall/runoff model is a lumped conceptual model that represents the surface water storage, the rootzone storage and the groundwater storage by three linear reservoirs. It is designed to represent accurately the water balance between the different storage zones in the soil. It is for example important to

remember that the "recharge" to the groundwater reservoir gives good information about the water balance but does not model accurately the dynamics of the actual recharge to the groundwater. In all, the model never provides more information than it has been designed for.

2.1.4 The estranged model

From observation to simplification of the processes of nature and calibration, the model becomes a different entity from the original system studied. It can represent most of the variations of the system studied but has also many limitations that the user tends to forget, especially with the upcoming of user-friendly models [1] that require little knowledge about the way the model actually works. It is therefore important to keep in mind the major sources of uncertainty when running a model. Melching in [15] provides a list of the major uncertainties encountered in hydrological modelling:

- Natural uncertainties: random temporal and spatial fluctuations. It is only possible to evaluate the magnitude of the uncertainties.
- Data uncertainties (forcing term): The main forcing term for hydrological models is the rainfall. The main problem comes from the discretization in time and space of the rainfall measurements and the necessity of spatial interpolation to get the rainfall over the whole domain of study from the information given by point measurements.
- Model-parameter uncertainties: It is impossible to find one set of parameters that represent reality properly.
- Model structure uncertainty: It is impossible to truly represent the physical processes by model simulation.

Additionally, scaling uncertainties must be taken into account. They are of three kind:

- The scale at which the differential equation represents the system. For example, the Richard's equation represents the behavior of the unsaturated zone at the scale of the centimeter;
- The scale of the model, i.e. the size of the discretization grid that is used to evaluate the equations;
- The scale of the observations that are used to build and calibrate the model. They can be point observations (piezometric heads or rainfall) or zone observations (remote sensing data).

It is essential to acknowledge the existence of these uncertainties in order to make good use of the models. The fact that these uncertainties exist is also a motivation to try to use the model not as the only source of information about the behavior of the system but in combination with the original observations. These observations have also a degree of uncertainty so techniques are needed to combine the different information.

2.2 From the Model to Nature: Data assimilation

2.2.1 A short definition of data assimilation

A good definition of data assimilation is given in [17]:

> The insertion of the reliable data into the dynamical model [...] to improve the quality and accuracy of the estimate.

The main motivation behind data assimilation is to try to add information to the model by using the observations. The observations have already been used during the calibration process of the model, but only to the extent of the concepts represented by the model. For example, if the model is linear, the calibration process leads to a set of parameters that represent best the linear part of the system modelled, and disregards the non-linear modes. In this case, data assimilation is an attempt to introduce some of the non-linearities.

The second interest in data assimilation is the possibility to add uncertainty estimation to the state estimation. Modelling and observation error are available to improve the modelling and observation strategy. As an example, monitoring networks for groundwater modelling can be improved to reduce modelling uncertainty using data assimilation [23].

2.2.2 Methodology

Hydrological simulation models can be referred to as process models [16]. They can be described as a set of equations that contain state variables and parameters. State variables vary with time whereas parameters remain constant (see Figure 2.2).

There are four methodologies for data assimilation:

- Updating of input variables: typically, precipitation and air temperature are updated. It is the classical method, justified by the fact that

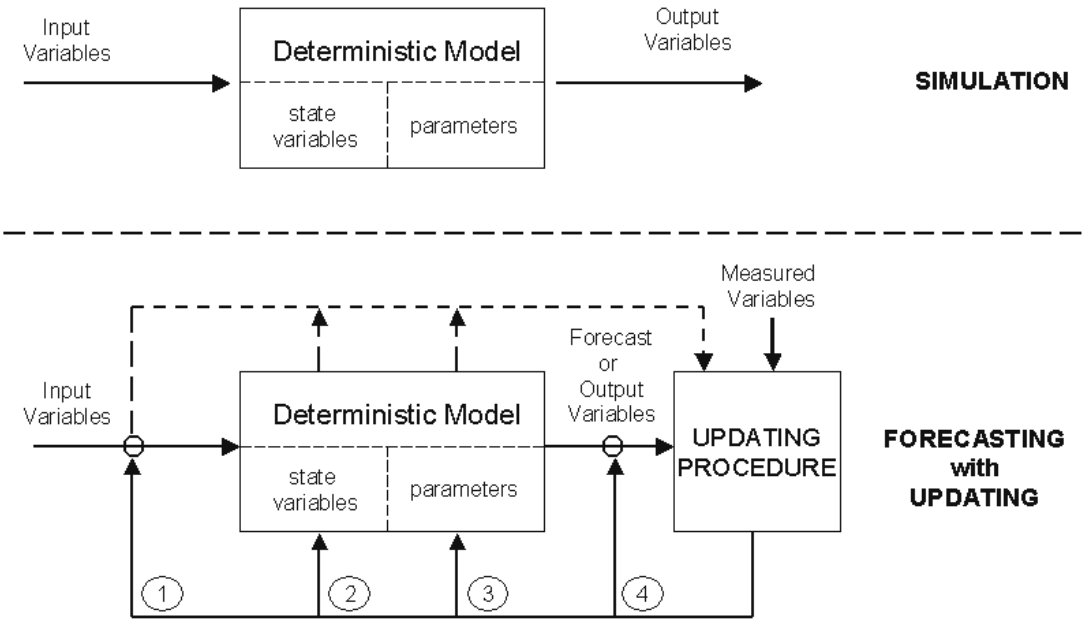

Figure 2.2: Comparison of the classical model run (upper part) with the model run with an updating procedure (lower part). Adapted from [16]

input uncertainty may often be dominant error source in operational forecasting.

- Updating of state variables: Adjustment of state variable (water content of conceptual reservoirs or piezometric heads in a distributed groundwater model for example) can be done in different ways. The most comprehensive theoretically based method is the Kalman filter.

- Updating of model parameters: the use of data assimilation to correct parameters is the least common as it is usually assumed that parameters do not vary in time.

- Updating of output variables (error correction): The deviation between the forecasted and the observed data are usually found to be serially correlated, making possible to forecast the future values of these errors by means of time series models like ARMA models.

2.2.3 Accepting partial ignorance: Stochastic representation

Stochastic model

Stochastic modelling is a way to accept the fact that the modelling process introduces uncertainties, as stated in Section 2.1.4. Even though there exist well developed mathematical theories about stochastic modelling (see for example [24]), most practical applications in hydrological modelling use simple error models that have a broader range of application.

The most simple stochastic model involves additive noise on a linear Markov chain model (compare with Equation 2.1):

$$\mathbf{x}_{k+1} = \mathbf{M}_k \mathbf{x}_k + \mathbf{u}_k + \eta_k \tag{2.9}$$

where η_k is a random variable whose probability distribution is defined according to the problem, referred to as *noise*. With a non-linear model M, the noise is generally incorporated as a part of the forcing [4]:

$$\mathbf{x}_{k+1} = M(\mathbf{x}_k, \mathbf{u}_k + \eta_k) \tag{2.10}$$

Because it is very difficult to decompose the noise models according to the different sources of uncertainties that have been described in Section 2.1.4, these basic models are the most successful in operational modelling: they provide a source of error that can be estimated either by rules of thumb according to the knowledge of the system or by the use of some adjustment methods ([6], **Paper C** and **Paper D**).

Stochastic observations

The uncertainty on the observations is modelled by:

$$\mathbf{y}_k^{\mathrm{o}} = H(\mathbf{x}_k^{\mathrm{t}}, \varepsilon_k) \tag{2.11}$$

where $\mathbf{y}_k^{\mathrm{o}}$ is the vector of observations at time k, $\mathbf{x}_k^{\mathrm{t}}$ is the true state of the system, ε_k the observation noise and H the observation operator that relates the state space to the observations. In practice, the true state of the system is not known and the observation equation is written as a function of the modelled state of the system $\mathbf{x}_k$ instead of the true state.

Ensemble approach to statistical modelling

Since it is generally not possible to manipulate directly the probability distributions, a popular approach (see [21] for example) is to model a sample population drawn from the desired probability distribution. Using the non-linear model of Equation 2.10, the ensemble approach works as follows:

1. Draw a population of m initial conditions $\mathbf{x}_0^1, ...\mathbf{x}_0^m$ from a given distribution

2. Draw an initial noise population of m noise vector $\eta_0^1, ...\eta_0^m$

3. Propagate the i^{th} member of the population using the model operator:

$$\mathbf{x}_1^i = M(\mathbf{x}_0^i, \mathbf{u}_0 + \eta_0^i) \tag{2.12}$$

4. For any time-step k, generate a noise population $\eta_k^1, ..., \eta_k^m$

5. Propagate:

$$\mathbf{x}_{k+1}^i = M(\mathbf{x}_k^i, \mathbf{u}_k + \eta_k^i) \tag{2.13}$$

This method is the most used and most generic method to account for any non-linearity in the propagation of the statistics of the state. It is the cornerstone of the Ensemble Kalman filter developed by Evensen [8, 7] discussed in more detail in **Paper A**.

2.2.4 Types of data assimilation

Data assimilation can be split into two different categories, according to the way the updating is done in time:

- Variational data assimilation: the past observations, from the start of the modelling until the present time, are used simultaneously to correct the initial conditions of the model and obtain the best overall fit of the state to the observations. This approach is essentially used in atmospheric sciences where the behavior of the system is driven by the accuracy of the initial conditions. The method is illustrated in Figure 2.3

- Sequential data assimilation: observations are used as soon as they are available to correct the present state of the model. In contrast to variational methods, sequential methods lead to discontinuities in the time series of the corrected state. This approach is more suitable in situations where the system is driven by boundary conditions. The method is illustrated in Figure 2.4

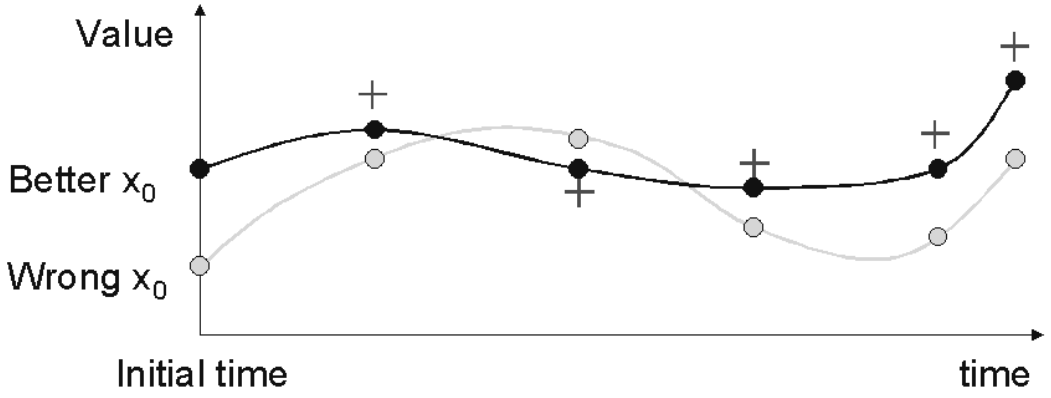

Figure 2.3: Variational data assimilation approach. The original model run (grey line and dots) is given better initial condition that leads to a new model run (black line and dots) that is closer to the observations(+).

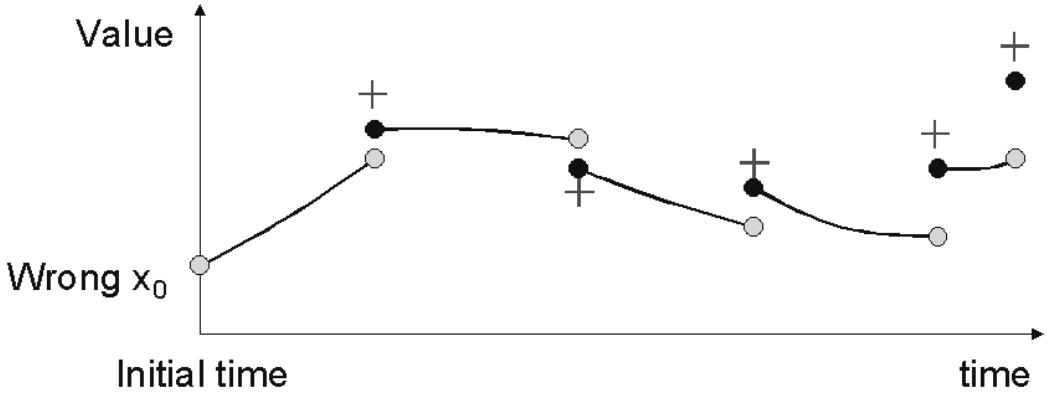

Figure 2.4: Sequential data assimilation approach. When an observation is available(+), the model forecast (grey dot) is updated to a value closer to the observation (black dot) that is used to make the next model forecast.

2.2.5 The Kalman filter and its assumptions

The Kalman filter is the most well-known data assimilation technique. Its popularity is due to the simplicity of its implementation and its relative robustness to the misspecification of the error sources [13]. The full mathematical derivation of the filter is available in **Paper A**. We will here discuss briefly the strengths and limitations of the filter.

The Kalman filter is a Best Linear Unbiased Estimate (BLUE). It means that it is optimal in situations where the model is linear. It is a minimum variance estimate, i.e. given an observation and a model forecast, it provides the estimate that minimizes the estimation variance. This property does not require any assumption about the distribution of the model error, just that the error is zero-mean uncorrelated in time. The meaning of the variance in case of non-Gaussian distributions, and especially in case of skewed distributions, needs to be specified in each case.

The advantage of using a Gaussian distribution is that the covariance of the state provides an estimate of the error between the state and the observations. Moreover this distribution is preserved by linear operators and is entirely defined by its two first moments. This is an important feature of the Kalman filter that is generally forgotten: it deals only with the two first moments of the distribution. Even the suboptimal schemes (for example the Extended Kalman filter or the Ensemble Kalman filter) that take into account the non-linearities of the model consider only the mean and the variance of the error during the updating procedure. It means that even though higher moments are propagated by the model, the way they are treated by the filter update is unpredictable. The simplicity of the Kalman filter is suitable in situations where the system is close to Gaussian and linear. In case of more complex distributions or non-linear models, the filter is still useful for the improvement of the performance but fails to provide an accurate uncertainty analysis.

In the case of groundwater modelling, and especially saturated zone modelling that has been the main issue of this work, the Kalman filter is well suited as the groundwater model is linear, and the assumption of having a unskewed distribution is most of the time respected. This condition is not respected anymore when the system is dominated by drainage. In this case, the maximum value of the piezometric head is constrained by the topographical elevation. This leads to a left-skewed distribution whereby the variance is not a good estimate of the error.

2.2.6 Data assimilation and the model

The main strength of data assimilation is the ability to extract the optimal amount of information out of the observations. Provided that the errors on the observations and the model are estimated properly, data assimilation gives the optimal estimate of the state of the system.

What makes the strength of data assimilation is also its main weakness. By forcing the model towards the observations, there is the possibility of introducing modes that the model cannot follow and therefore drive the model out of its stability domain. This is especially true in the case of sequential data assimilation where the model variations are seen only at a given time step. The dynamics of the model do not influence directly the result of the analysis.

It is therefore important to validate the assumptions of the errors to avoid too strong a forcing. **Paper C** provides a technique whereby the errors are actually calibrated to avoid this problem.

Using data assimilation also puts the model into a new perspective. Without data assimilation, the model is the only source of information about the system studied, but the use of observations at assimilation points turns the model into a kind of complex extrapolation operator. If it becomes so, there is a risk of using the physical model as a black box. The calibration results of **Paper D** are typical examples of this, as some of the resulting parameters have little physical meaning but still lead to the best model performance.

Chapter 3

Papers and research strategy

3.1 Research path and choices

The overall goal of this thesis has been to try to put the optimization of the Kalman filter and its uncertainty analysis properties into a coherent framework for specific application in groundwater modelling. However the developed procedures are generic in nature and hence applicable to other modelling domains.

During the study, it became clear that bias was an important problem and needed to be taken into account by the filter itself in order to get relevant results. This resulted in the study and improvement of the bias aware methods presented in **Paper B**.

The attempt to identify the different sources of uncertainty and apply them in the framework of the ensemble Kalman filter proved to be a difficult endeavor. Using literature to quantify uncertainties on parameters, forcing and model, did not lead to any meaningful result. It motivated the return to a very simple additive error model that could be manipulated and tuned easily. The use of an additive error was therefore a conscious choice resulting from the failure of more complex models.

Even the tuning of this simple model was a problem that was little addressed in literature, especially when linked to groundwater modelling. The use of automatic calibration instead of rules of thumb came as a natural approach: if the user is able to tune the noise parameters, why not use an objective algorithm instead? Such a calibration procedure was developed and is reported in **Paper C**.

Using Kalman filtering in combination with automatic calibration without controlling the stability of the model outside of the assimilation points would automatically lead to overfitting. Therefore the model behavior objective

function Δ_x was devised. The original measure was aimed at controlling the behavior of the state filter and the bias filter. The state filter innovation should be kept with a mean of zero while the bias was corrected by the bias filter. The final form of the measure, i.e. the measure of model variation between two time steps, is a classical case of serendipity. While building the model, the author defined wrong data flows and realized that the resulting measure was actually more effective than the intended one.

The developed sequential calibration method was discovered through the analysis of calibration results where the observation variance was one of the calibration parameters. The results showed that very different Kalman filter setup lead to very similar performances. This lead to the attempt and success to prove theoretically that, in the conditions of the experiments, the Kalman filter results did not depend directly on the observation variance.

Finally, the calibration method was applied to a real case, the Danish Karup catchment. The model and the filter parameters were calibrated simultaneously to test the influence of the data assimilation on the parameter estimation. The results are analyzed in **Paper D**.

3.2 Overview of the papers

Paper A is a literature review of the Kalman filter techniques and their application in hydrological modelling. It gives the theoretical basis for the research undertaken during this work.

Paper B is a theoretical paper that compares two different bias aware Kalman filter technique. On the one hand, the Separate Bias Kalman filter that assumes no correlation between the bias and the unbiased state, and on the other hand a more general technique, the colored noise Kalman filter. The use of bias feedback into the model forecast shows to be more flexible than modelling the bias aside. The conclusions of this paper have been applied to the two following papers, **Paper C** and **Paper D**.

Paper C presents the central achievement of this thesis work: the sequential calibration technique that leads to a better performance of the Kalman filtering technique and also an uncertainty analysis based on hard data instead of rules of thumb. The paper provides the analysis of the results of the calibration method using a twin-test experiment. The use of a twin-test experiment allows for the comparison between the estimated results and the truth. The paper demonstrates the strength of the technique but also points out the problems of underestimation of the uncertainty when information is scarce.

In **Paper C**, the Kalman filter parameters are calibrated alone and it

leads to problems in the representation of the dynamics of the system. In **Paper D** it is attempted to better account for the dynamics of the system by calibrating both the model and the Kalman filter parameters simultaneously. Using the technique proposed in the paper, the results are not entirely satisfying as the physical model is used only for interpolation and does not provide any insight about the system studied.

Chapter 4

Conclusions, discussion & future work

The work in this thesis intended to circumvent essential problems when dealing with automatic calibration and uncertainty analysis combined with data assimilation:

- The use of rules of thumb to estimate the parameters of the covariance matrices in the Kalman filter implementation;
- The risk of overfitting the observations by overusing data assimilation.

As it is mentioned in the opening quote, it is essential not to confuse exactness and truth. By using data assimilation, it is possible to match exactly the values of the observations, regardless of the behavior of the model over the whole domain. The lack of care in applying an automatic calibration technique in collaboration with data assimilation can lead to a model that does not provide insight about the system studied but let data assimilation give the illusion of an exact match, yet without any relation to the truth.

However, by carefully tuning the covariance parameters of the data assimilation technique, it is possible to acknowledge the imperfection of the model and the uncertainty on the observations. Automatic calibration becomes a tool to adjust the different parameters describing the model and its inherent uncertainty. Since uncertainty is accounted for, the physical parameters of the model are the best parameters possible, not in the sense of an exact match, but because they lead to the best representation of the true system, given the assumptions that were used to build the model.

The results of **Paper D** show that the work done during this thesis is on the track of achieving such goals. Nevertheless the results need to be refined

in order to achieve better physical insight about the systems modelled. Here are listed some improvements that are likely to be beneficial:

- The calibration of the system {data assimilation + physical model} needs to be refined. The calibration algorithm needs to "see" the physical model as such, and not as a hidden part of the data assimilation system. As an improvement of **Paper D**, it is suggested to use two calibration periods, one with data assimilation and one without and use objectives functions from the two periods in a Pareto optimization framework to study the tradeoff between data assimilation and physical modelling;

- The physical knowledge of the system needs to be incorporated into the covariance model, either in the form of a different correlation structure (e.g. by using information about the correlogram) or in the form of a correlation distance that varies according to the location;

- The physical model has an influence on the ensemble generation and the uncertainty analysis. This influence needs to be studied in more detail, especially in cases when the model leads to skewed ensemble distributions that are not properly dealt with by the Kalman filter;

- The feedback of the bias, which had a beneficial influence on the model in artificial cases, could cause problems in real-world situations. One of the main reasons for this is the problem of representativeness of the observations: grid based variables of the model are compared with point based observations that are not necessarily representative of the modelled grid value. This includes an additional bias that is only present at assimilation points. It would be interesting to estimate this additional bias and correct it before using the observations for assimilation;

- The persistent model used as bias forecast model is valid as long as observations are available. But during a forecast period, it would be useful to use a more complex model that takes into account the decreasing importance of the old observations compared to the model.

Bibliography

[1] V. Babovic. *Emergence, Evolution, Intelligence: Hydroinformatics*. A.A. Balkema, Netherlands, 1996.

[2] K. Beven and A. Binley. The future of distributed models: Model calibration and uncertainty prediction. *Hydrological processes*, 6:279–298, 1992.

[3] K. H. Brink and A. R. Robinson, editors. *The Sea, Volume 10*. John Wiley & Sons, 1998.

[4] R. Cañizares, H. Madsen, H. R. Jensen, and H. J. Vested. Developments in operational shelf sea modelling in Danish waters. *Estuarine and Coastal Shelf Science*, 53:595–605, 2001.

[5] V. T. Chow, D. R. Maidment, and L. W. Mays. *Applied Hydrology*. McGraw–Hill, 1988.

[6] D. P. Dee. On-line estimation of error covariance parameters for atmospheric data assimilation. *Monthly Weather Review*, 123:1128–1145, 1995.

[7] G. Evensen. Sequential data assimilation with a non-linear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research*, 99(C5):10143–10162, 1994.

[8] G. Evensen. The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dynamics*, 53(4):343–367, 2003.

[9] J. Freer, K. Beven, and B. Ambroise. Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach. *Water Resources Research*, (7):2161–2173, 1996.

[10] H. V. Gupta, S. Sorooshian, and P. O. Yapo. Towards improved calibration of hydrological models: Multiple and noncommensurable measures of information. *Water Resources Research*, 34(4):751–763, 1998.

[11] K. Havnø, M. N. Madsen, and J. Dørge. *MIKE 11 – a Generalized River Modelling Package*, chapter 21, pages 733 – 782. In Singh [20], 1995.

[12] M. A. Hooimeijer. *Reduction of Complex Computational Models*. PhD thesis, Delft University of Technology, 2001.

[13] H. Madsen and R. Cañizares. Comparison of extended and ensemble Kalman fitlers for data assimilation in coastal area modelling. *International Journal For Numerical Methods in Fluids*, 31:961–981, 1999.

[14] A. Mees, editor. *Nonlinear Dynamics and Statistics*. Birkhauser, 2000.

[15] C. S. Melching. *Reliability Estimation*, chapter 3, pages 69–118. In Singh [20], 1995.

[16] J. C. Refsgaard. Validation and intercomparison of different updating procedures for real time forecasting. *Nordic Hydrology*, 28:65–84, 1997.

[17] A. R. Robinson, P. F. J. Lermusiaux, and N. Q. Sloan III. *Data Assimilation*, chapter 20, pages 541–593. In Brink and Robinson [3], 1998.

[18] F. Rubel and M. Hantel. Correction of daily raingauge measurements in the Baltic sea drainage basin. *Nordic Hydrology*, 30:191–208, 1999.

[19] J. R. Schewchuck. An introduction to the conjugate gradient method without the agonizing pain. Technical report, School of Computer Science, Carnegie Mellon University, 1994.

[20] V. P. Singh, editor. *Computer Models of Watershed Hydrology*. Water Resources Publication, 1995.

[21] L. A. Smith. *Disentangling Uncertainty and Error: On the Predictability of Nonlinear Systems*, chapter 2, pages 31–64. In Mees [14], 2000.

[22] F. Vejen, P. Allerup, and H. Madsen. Korrection for fejlkilder af daglige nedbørmålinger i Danmark. Technical Report 98-9, Danish Meteorological Institute, 1998.

[23] Z. Yangxiao, C. B. M Te Stroet, and F. C. Van Geer. Using Kalman filtering to improve and quantify the uncertainty of numerical groundwater simulations. 2. Application to monitoring network design. *Water Resources Research*, 27(8):1995–2006, 1991.

[24] D. Zhang. *Stochastic Methods for Flow in Porous Media*. Academic Press, 2002.

THE DAIHM TOOLBOX

Part of the work done during the Ph.D. involved the development of two versions of a Matlab® toolbox. The motivation was the lack of tools to manipulate time series in the form of ensembles, as well as the transition between a state representation of a model (needed during the data assimilation procedure) and a spatial representation (needed for the physical modelling). The first version (v1.02) of the toolbox simply aims at organizing the communication between different functions and the use of an ensemble of time series. The second version (v2.0) is more advanced and has been used to run the models presented in the research papers. This section gives an overview of the features of the toolbox.

State-Space representation

The necessity to use a state representation of data that are spatially distributed, i.e. to turn data that are best represented on a 2D or 3D grid (the *space representation*) into a one-dimensional object (the *state representation*) is an important issue of data assimilation. In case of rectangular or parallelepipedic domains, the transformation is straight-forward. For irregular domains like hydrological catchments, we need to be certain that the transformation from space to state and vice-versa is unique and will always lead to the same domain and the same state.

The concept of *mapping* is introduced. The mapping array is an array that has the same shape as the spatial array that contains the data. Where no data exist, the value `NaN` is used (Not A Number). Where data are available, the mapping array contains the index of the line in the vector of the state representation. This means that the mapping array gives information of where to place in the state a given variable at a given location in the space representation, and also where to put in the space representation a given variable in the state representation.

Efficient functions have been written to allow transformation between the state and the space representation using the DAIHM data format.

The DAIHM data format

Given the ability to change between state and space representation, it is necessary to store the time series and ensemble data together with the mapping information in a single variable that can be manipulated as a whole.

A variable of the DAIHM data format is a structure composed of the following fields:

- `Name`: String. Indicates the content of the variable;
- `Param`: Boolean. It is given the True value in case the variable is a parameter, i.e. it does not depend on time. This field is needed for the storage of data on disk.
- `EnsSize`: Integer. The size of the ensemble of the variable (See field `State`). If the size is 0, the variable is a *statistical* variable, i.e. it represents a statistical property of an ensemble (for example a covariance matrix);
- `Intensive`: Boolean. Indicates whether the interpolation between two time steps should be taken as a linear interpolation (True) or divide the given value in smaller quantities (False). This is useful for rainfall data for example. If it rains 10 mm in one day, and one wants to know the rain by increments of 12 hours, the result should be 5 mm during the first 12 hours and 5 mm during the following 12 hours.
- `Units`: Structure. A field describing the units as exponents of elementary units. It is used to ensure good compatibility between the models unit input and the variables units.
  - `Abrev`: String. The SI designation of the unit;
  - `LMT`: 1×7 Array of Double. Gives the exponent of the 7 elementary units: meter, kilogram, second, Ampere, Kelvin, moles, candela. For example a `LMT` equal to `[1 0 -2 0 0 0 0]` means an acceleration in m/s^2;
  - `Fact`: 1×7 Array of Double. Gives the scale factor to each of the units. For all the units except the seconds, it is given in the form of the decimal logarithm. For example a `Fact` of -3 for the meter unit means that the system deals with mm. For the temporal unit, the natural exponents are used: ms, s, h, d, y.
  - `Name`: String. A field used to categorize the unit, such as distance, acceleration, temperature.

- `Mapping`: Array of integers. The mapping array of the variable for state-space transformation. All the fields mentioned until now make the header of the variable. The last two fields contain the data themselves.

- `TSteps`: Double. The time steps of the variable where values are available. The time steps are recorded as absolute values using the internal date indexing of Matlab. If the variable is a parameter, the field is empty.

- `State`: Double. The data in state representation. The `State` field is generally a 3D array. The first dimension (rows) is used for the state representation, the second dimension (columns) is used to store the different members of an ensemble and in the third dimension (depth), the different time steps are stacked. For a variable that is not a parameter, the depth of the `State` field is the same as the length of the `TSteps` fields. This approach has different advantages: (1) An ensemble at a given time step is a matrix that can be processed all at once if the model is a matrix itself; (2) In case of a parameter, the field is at most 2D, there is no need to manipulate clumsy 3D arrays; (3) It facilitates the storage of data on disk.

The structure of the DAIHM data format is motivated by the need to store data on disk time step by time step as the model runs. It is possible to create the header as a preprocessing and then store the data incrementally as needed. The final file, with the extension `.ddat` (DAIHM data), is a binary file containing the header information that can be loaded separately and then the time series data in chronological order associated with their time tag.

The model building utility

The Graphical User Interface of the model building utility is shown in Figure 1. It helps the user setting up a complex model from simple functions, while fully using the features of the DAIHM data structure. The underlying concept of this utility is the use of the Markov Chain model, i.e. the model uses the previous time step information, some forcing and parameters to forecast the current time step. An example of a resulting model is shown in Figure 2.

The model building utility is based on two concepts:

- The model, represented as a box on the flowchart. It encapsulates information about the input/output data necessary for a given function

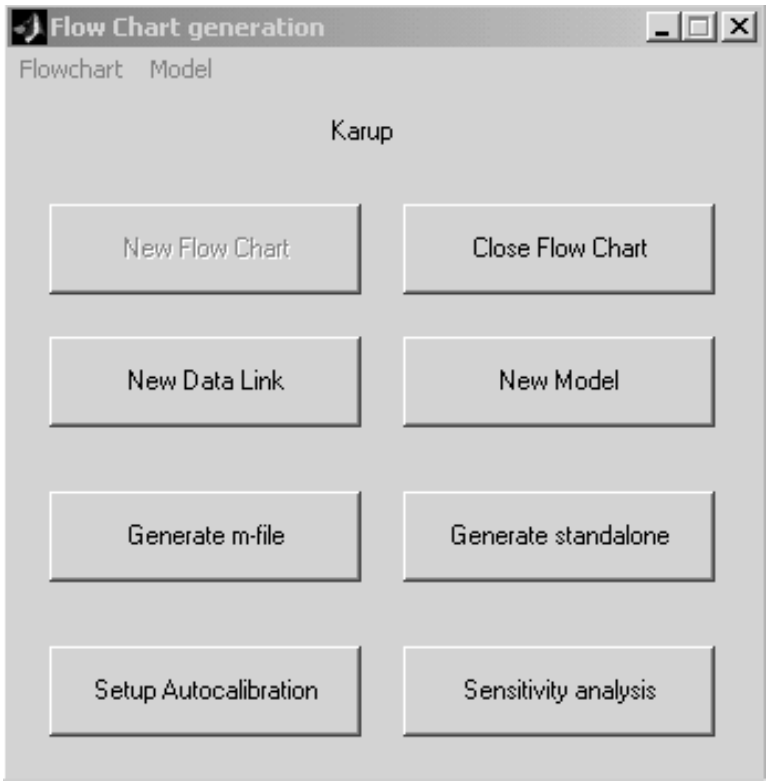

Figure 1: Screen capture of the Graphical User Interface of the DAIHM toolbox.

to be run. This information is provided by a dialog box (Figure 3) that defines for example the unit expected by the model or the representation of the variable (state or space). Each model can be involved either in the preprocessing, the processing or the postprocessing.

- The data streams, represented by the arrows on the flowchart. They define how data is transferred from one model to the other.

The model building utility ensures that the flow of data between two model boxes is consistent and that units, ensemble size and state/space form are respected. It also defines the input, parameters, output of the model, including the storage frequency.

When the model is drawn, an m-file is generated automatically that allows to run the full model from the Matlab command prompt. The generation of the m-file ensures that the model setup is coherent, that no input or output have been forgotten, and especially that any cycle in the flow of data passes through the special model `NextTS` that has the only function to propagate the variable values from one time step to the next. The final m-file takes into

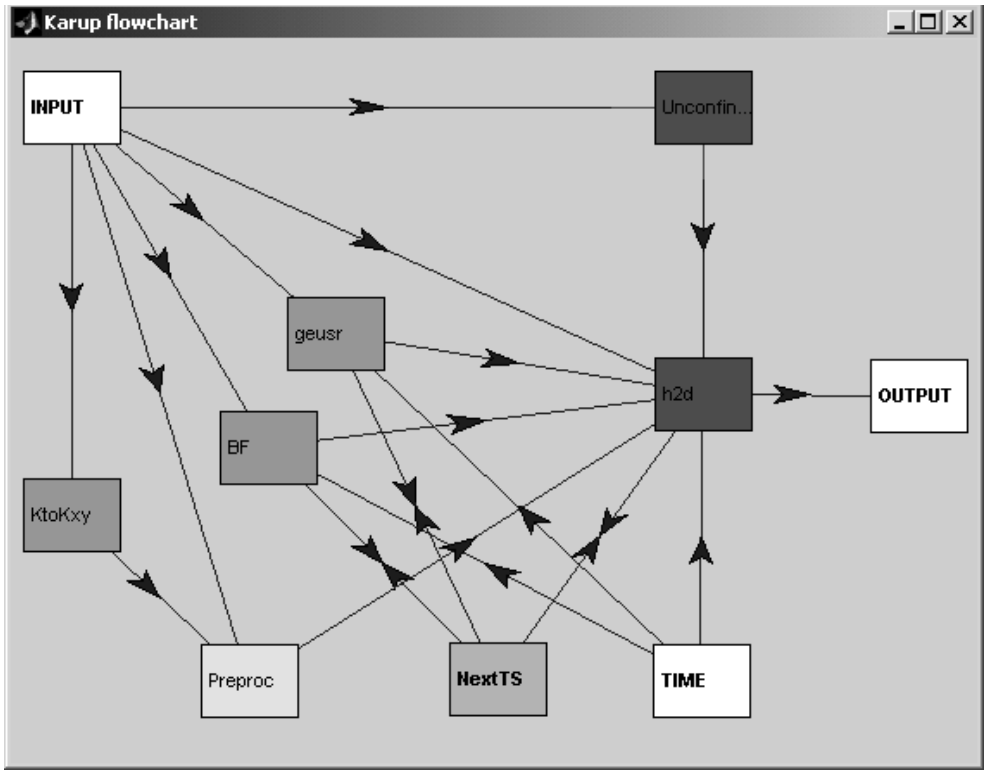

Figure 2: Screen capture of the flowchart of the groundwater model used in the project.

account the pre- and postprocessing of the data and manages the storage of the output in a proper form. An additional text file is generated to allow the user to change the input/output options as well as the time step and the simulation time without the need to recompile the function.

The m-file, its parameter file and the data necessary to run the model are stored into a folder. An additional feature allows to compile the resulting model into a DOS-executable (using the Matlab `mcc` function). The executable, included into the model folder, can be run on any PC, provided that the free Matlab library has been installed. It is therefore possible to run the resulting code on a machine that does not have Matlab installed. In the present version of the toolbox, the compilation as well as the processing of the `.ddat` files needs to be done under the Matlab environment.

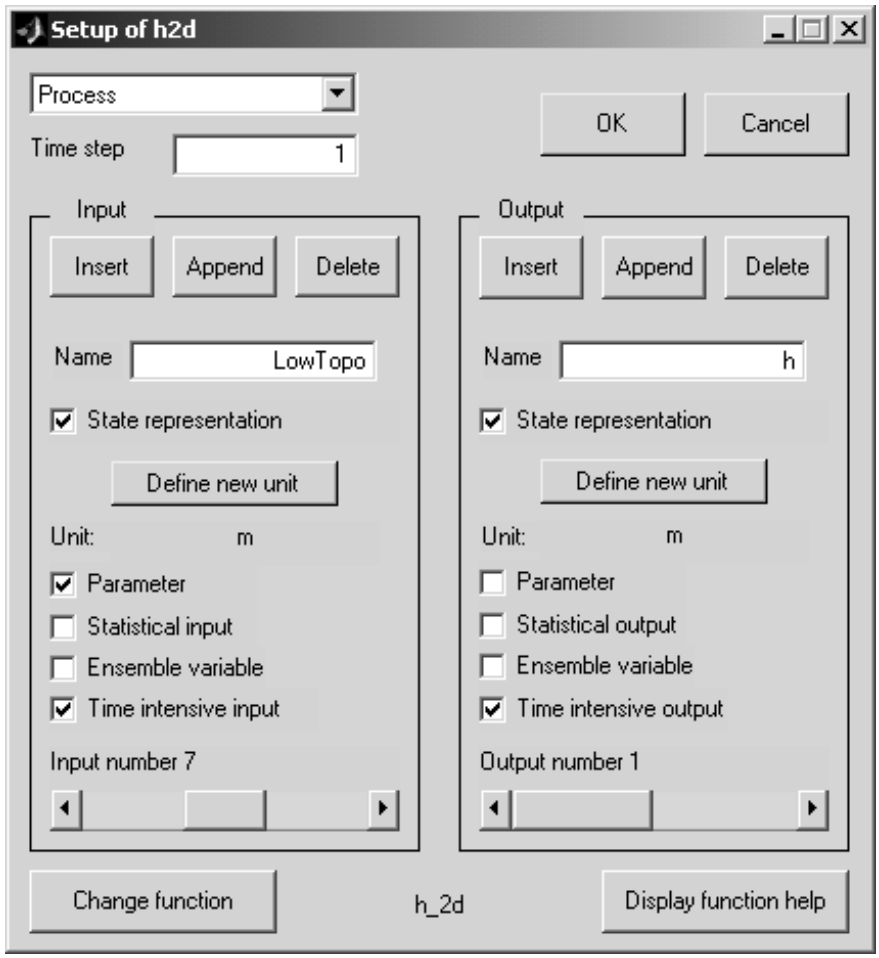

Figure 3: Screen capture of the GUI for the setup of the input/output options

Additional tools

Morris sensitivity analysis

The model can be automatically setup to run a Morris OAT sensitivity analysis (cf. **Paper D**). The results are gathered in a specific sub-folder of the model folder and can be easily plotted into the graphical interpretation of the results.

When specified by the user, the sensitivity analysis can be run as an ensemble simulation, i.e. the different points constituting the random trajectory necessary for the analysis are run simultaneously. If the functions that are used in the model are adequately programmed, it can lead to a significant improvement in speed as the assignment of constants and the preprocessing/postprocessing are done only once.

Autocalibration software interface

The setup of the automatic calibration tool, AUTOCAL developed by at DHI Water & Environment and used in **Paper C** and **Paper D**, has been made compatible with the `.ddat` format. The setup file necessary to run the automatic calibration is generated automatically by the model building utility and the DOS-executable version of the model is used to run the calibration. The results are gathered in a specific folder. The main feature of this utility is the possibility to transfer data from `.ddat` files to text files and vice-versa.