Improving hydrogeological models using the results of calibrated groundwater flow models

A probabilistic approach using piecewise linear probability density functions and Bayesian networks

Aris Lourens

229 UTRECHT STUDIES IN EARTH SCIENCES

Improving hydrogeological models using the results of calibrated groundwater flow models

A probabilistic approach using piecewise linear probability density functions and Bayesian networks

Verbeteren van hydrogeologische modellen door resultaten van gecalibreerde grondwatermodellen te gebruiken

Een probalistische benadering met behulp van gelineairiseerde kansdichtheidsfuncties en Bayesiaanse netwerken

(met een samenvatting in het Nederlands)

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof. dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

vrijdag 16 april 2021 des middags te 2.30 uur

door

Aris Lourens

geboren op 21 januari 1962 te Aalten

Promotoren:
Prof. dr. ir. F. C. van Geer
Prof. dr. ir. M. F. P. Bierkens

Utrecht Studies In Earth Sciences 229

Improving hydrogeological models using the results of calibrated groundwater flow models

A probabilistic approach using piecewise linear probability density functions and Bayesian networks

Aris Lourens

Utrecht 2021

Promotoren:
Prof. dr. ir. F. C. van Geer
Prof. dr. ir. M. F. P. Bierkens

Examination committee:
Prof. dr. H.J. Hendricks-Franssen
Prof. dr. J.C. Refgaards
Prof. dr. ir. A.W. Heemink
Prof. dr. ir. M. Bakker
Dr. S.B. Dubinkina

This work was financially supported by TNO - Geological Survey of the Netherlands.

ISSN 2211-4335

ISBN 978-90-6266-594-5

Cover: Four scientists whose work was a foundation for this thesis. From left to right: Thomas Bayes, Danie Krige, Henry Darcy and Winand Staring. Depicted in the scene of Landmannalaugar, Iceland.
Painted by Vera Lourens.

Print: Ridderprint | www.ridderprint.nl.

Improving hydrogeological models using the results of calibrated groundwater flow models

A probabilistic approach using piecewise linear probability density functions and Bayesian networks

Het kan gebeuren dat je uren zit te denken en er schiet je maar niks te binnen,
en zo maar in een keer bereik je precies hetzelfde in nog geen vijf minuten.

Herman Finkers
uit: Geen spatader veranderd

Voorwoord

WAAROM ZOU JE eigenlijk promotieonderzoek doen? Zoals Bomans[1] al zei: "De kern van deze vraag zit natuurlijk in het woordje 'eigenlijk'. Men zou er beter aan doen het te laten." Maar toch, op die vraag zijn veel goede antwoorden te geven waaruit blijkt dat je het eigenlijk wèl zou willen doen.

Mijn belangrijkste goede antwoord is: omdat het leuk is! Dat wil zeggen, voordat je er aan begint heb je sterk dat vermoeden. In mijn geval blijkt dat vermoeden meer dan gegrond. Het was inderdaad heel erg leuk. Zo leuk zelfs dat ik er maar niet mee kon stoppen, bijna tien jaar lang.

Een tweede goed antwoord is: omdat je de kans krijgt. Frans en Marc, ik ben jullie dankbaar dat jullie mij die kans hebben gegeven. Of dat misschien *fingers crossed* was hebben jullie me nooit laten merken. En Deltares heeft mij de kans gegeven door mij een terugkeergarantie te geven. Dat gaf toch wat prettige zekerheid en maakte de stap gemakkelijker. Want ja, een leuke baan opzeggen voor 'vier' jaar onderzoek en niet te weten wat daarna komt...

Nog een goed antwoord: je leert er veel van. Dat is zeker waar, er is echter een keerzijde heb ik gemerkt. Want bij alle kennis die je vergaart blijkt dat er nog veel meer is dat je niet weet. En dat loopt behoorlijk op in een paar jaar tijd, kan ik wel zeggen. Het is misschien leuk om dit proces als een Bayesiaans model te zien, daar gaat het in dit proefschrift tenslotte ook over. Bij een Bayesiaans model mag je een beginschatting maken van de situatie. Dat doe je met de kennis die je tot dan toe hebt. Vervolgens voeg je kennis toe en kom je tot een aanpassing (update) van je eerste inschatting. Stel nou dat ik aan het begin van mijn promotieonderzoek dacht dat ik ongeveer de helft van de bestaande wiskunde wel zo'n beetje beheerste of zou begrijpen. Zie het als jeugdige overmoed. Na jaren van bestuderen van allerlei wiskundige onderwerpen heb ik hier heel wat kennis aan toegevoegd. Voeg deze kennis toe aan het Bayesiaans model en wat blijkt, ik zit nu op ongeveer één procent! En ik vrees het ergste voor als ik verder studeer.

Kom, ik gooi er nog een goed antwoord tegenaan: omdat je het niet alleen hoeft te doen. Dat antwoord lijkt wat vreemd, het is juist de bedoeling dat je laat zien dat je zelfstandig onderzoek kunt doen. Maar toch, hulp is onontbeerlijk. Uiteraard, je promotoren denken mee in welke richting het onderzoek moet gaan, welke ideeën uitgewerkt kunnen worden en wat je misschien maar beter kunt laten. Maar daarnaast kun je nog wel wat hulp gebruiken. Je hebt bijvoorbeeld data nodig, veel. In mijn onderzoek ging het om gegevens van de ondergrond (het REGIS model) en om

[1] vrij naar Godfried Bomans, uit: Waarom schaakt u eigenlijk?

gegevens van een grondwatermodel (het AZURE model). Voor gegevens van het eerste model hebben vooral Eppie de Heer en Jan Hummelman mij enorm geholpen en voor het tweede model Joachim Hunink. Naast gegevens ga je ook gebruik maken van nieuwe gereedschappen, ook daar kun je wel wat hulp bij gebruiken. Alle dataverwerking in FORTRAN oplossen leek toch wat omslachtig, een taal als R blijkt dan een zeer goede aanvulling. Gelijk aan het begin van mijn onderzoek wilde ik wat gedachten opschrijven en daar kwamen wat formules in voor. Vervolgens werd ik binnen een week gillend gek van Word. Dus rende ik naar Niko Wanders of Edwin Sutanudjaja en kreeg hulp bij R en LaTeX. Wat een verademing. En dan zijn er natuurlijk nog veel meer mensen die met van alles en nog wat geholpen hebben. Zonder al deze hulp was het niks geworden.

Vooruit dan, nog een goed antwoord: om een schuld te vereffenen. Ik heb het voorrecht gehad om paranimf te zijn bij de promoties van Wilbert Berendrecht en Peter Vermeulen, samen met Peter bij Wilbert en samen met Wilbert bij Peter. Jongens, ik ben blij dat jullie het geduld hebben kunnen opbrengen en nu bij mij als paranimf op willen treden. Het figuur blijkt dan toch weer een vorm te hebben waarbij de afstand tot het middelpunt overal gelijk is. Zo kunnen we het mooi afsluiten.

Als ik dan toch bezig ben, een goed antwoord is ook: omdat je veel aandacht krijgt. Oh, wil je dan graag in het zonnetje gezet worden? Nou nee, dat was niet de insteek maar je krijgt het wel cadeau. Heel veel mensen vroegen belangstellend hoe het met mijn onderzoek ging. Na verloop van tijd werd die vraag zelfs wat beschroomd gesteld: "ik weet niet of ik het mag vragen, maar..." Veel vrienden, kennissen, collega's (of hier een combinatie van) toonden zeer regelmatig belangstelling, en dat is heel leuk. In het bijzonder wil ik hier mijn familie noemen. In de eerste plaats mijn ouders. Ik ben blij dat jullie mijn strapatsen met zoveel interesse hebben gevolgd. En natuurlijk mijn schoonouders, mijn broer en zussen, mijn schoonzussen, zwagers, ooms en tantes. Van zoveel belangstelling ga je bijna naast je schoenen lopen.

Als afsluiting het allermooiste goede antwoord: Linda! Lieve Linda, dank je wel dat ik dit mocht doen en dank je wel voor al je steun. En natuurlijk Vera, Maarten, Daniel en Caroline. Ook jullie hebben het mogelijk gemaakt dat ik dit kon doen. Het is niet bepaald aan jullie voorbij gegaan dat er heel wat uurtjes in dit boekje zijn gaan zitten. Jullie zijn geweldig. En voor wie het ontgaan is, Vera heeft het prachtige schilderij op de voorkant gemaakt.

Ik wens iedereen veel leesplezier.

Contents

1

General introduction

MODELS are generic tools to simplify real world problems. They don't solve the problems but are used to approximate unknown representations of reality to make the real world problems tangible and to assist decision making. Models never describe reality completely but only give a more or less rough impression of it. If a model is well defined, it may be used to obtain quantitative estimates for some quantities it describes. Examples of applications of models are found in meteorology, where models are used for the daily weather forecast, in economics, where models are used to predict the change of the unemployment rate caused by a certain tax measure, in (geo)hydrology, where models are applied to predict the river stages or groundwater heads caused by precipitation, and in exploration geology, where models are used to infer the likelihood of finding natural resources like ore and oil. Also combinations of these models exist, when, for instance, the predicted precipitation of the meteorological models are used to predict the river stages with hydrological models for the oncoming days.

All these models are based on assumptions and simplifications and have in common that they only can make predictions for certain parameters within some degree of uncertainty. It is always a challenge to the modeler to decrease the uncertainty of a model and improve its predictive power. The performance of a model can be tested against the available observations of the predicted variables. For instance, the river stages for the oncoming days can be predicted by using the weather forecast, while after the predictions have been made, observed river stages are compared to the predictions. The differences between the observations and the model predictions are a measure of the accuracy of the model, and in general the smaller the deviations the better the performance of the model. The deviation between predictions and observations are often used to improve the model by calibration: changing the model setup, usually its parameters, such that the deviation is minimized. So the use of observations of the modeled process can improve the model performance.

This thesis is concerned with two connected models (Figure 1.1). The first is a hydrogeological model which describes the subsurface in terms of layers with relatively high hydraulic conductivity materials (aquifers) and layers with low hydraulic conductivity materials (aquitards), and the accompanying parameterization of the layer thickness and the conductivity. The second model is a groundwater flow model which mimics the dynamic behavior of the groundwater in the subsur-

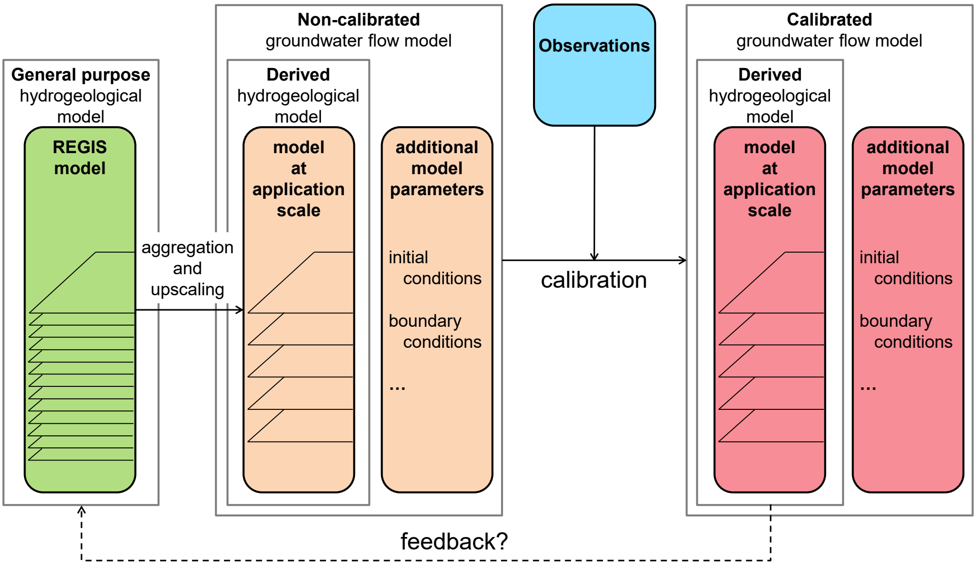

Figure 1.1: *General problem definition. The hydrogeological model for a groundwater flow model is derived from a general purpose hydrogeological model, (in this thesis the REGIS model). Updates during calibration do not affect the general purpose hydrogeological model. A feedback procedure is developed in this thesis.*

face. The subsurface description of the groundwater flow model is derived from the hydrogeological model, which creates a connection between the two models. This derivation usually involves aggregating hydrogeological layers with high conductivities into single aquifers and estimating the aquifer transmissivities (m^2/day) by upscaling from the conductivities of the layers being aggregated. Similarly, aquitards are defined by aggregating low-conductivity layers and estimating the aquifer resistivity or C-value (days) from upscaling. The resulting groundwater flow model, that also includes boundary and initial conditions, is usually calibrated, including its hydrogeological parameters, using, for instance, observations of groundwater heads of the modeled area. Herewith, the groundwater flow model, including its derived hydrogeological model, is calibrated but the underlying hydrogeological model is not. This causes an inconsistency between the groundwater flow model and the underlying hydrogeological model that is generally not resolved. The reason that it is not resolved in practice is that there is no unique one-to-one relationship between the calibrated transmissivities and resistivities that result from groundwater model calibration and the original conductivities and layer thicknesses of the underlying hydrogeological model. Yet, even under conditions of non-uniqueness and uncertainty, it must be possible to develop a method to improve the underlying hydrogeological model using groundwater model calibration results. This is the main objective of this thesis.

1.1 Hydrogeological model of the Netherlands

At the Geological Survey of the Netherlands (TNO-GSN), a three-dimensional digital geological model (DGM) [*Gunnink et al.*, 2013] has been developed and is continuously maintained. With this model, the subsurface is subdivided in geological units. The definition of the units is based on the age of the deposits and on the depositional environment, like marine or fluvial. Most units coincide with a geological formation.

The DGM serves as a framework for multiple models, which make a refinement within the geological units. An example of these models is the GeoTOP model [*Stafleu et al.*, 2011], a voxel model which describes hydrogeological parameters of the subsurface up to a depth of about 50 meter. Each voxel of $100\,\mathrm{m} \times 100\,\mathrm{m}$ and $0.5\,\mathrm{m}$ thick is parameterized with hydrogeological and lithological data. Another model is the hydrogeological model REGIS-II [*Vernes et al.*, 2005; *Vernes and van Doorn*, 2006]. This model defines the hydrogeological units (typically layers of varying thickness and hydraulic conductivity) up to a depth of about 500 meter. This REGIS model is used in this thesis. The definitions of the hydrogeological units are based on the layers' hydraulic properties, mainly the hydraulic conductivity. Each unit is typically modeled as a high resistance layer (aquitard) or low resistance layer (aquifer). In this model, over one-hundred hydrogeological units are recognized. Since REGIS is defined for the whole of the Netherlands (European mainland), not all units are present in each area. To keep the model consistent, each unit is defined everywhere in the model domain, but locally absent units are locally modeled with zero thickness. REGIS is not developed for one specific purpose but serves as a generic model for all studies that need a consistent description of the hydrogeologic and hydraulic properties of the subsurface of the Netherlands.

Many groundwater models serve a specific purpose and are therefore developed locally and not for the country as a whole. Therefore, for each separate groundwater modeling study a submodel (the derived hydrogeological model) is derived from REGIS, which is further processed to meet the needs of that specific study. Such processing often aggregates multiple REGIS units into one aquifer or aquitard. Herewith, the REGIS model and the derived hydrogeological model are equivalent but not equal (Figure 1.1).

As stated, all models are uncertain[1], and so are the REGIS data. Therefore, a derived hydrogeological model which serves as a hydrogeological model for a groundwater flow model is usually calibrated. Since the derived model is separated from the main model REGIS, the calibration of the model parameters do not affect the REGIS model. To let REGIS benefit from the calibration, a feedback procedure is needed to update the REGIS parameterization too (Figure 1.1). This update is not directly beneficial for the readily calibrated groundwater flow model, but it

[1] We should in fact say that model outcomes, model parameters and data are subject to uncertainty, i.e. we as humans are uncertain about their real values. However, in this thesis, this formulation is shortened to 'uncertain parameters, data or models' for convenience.

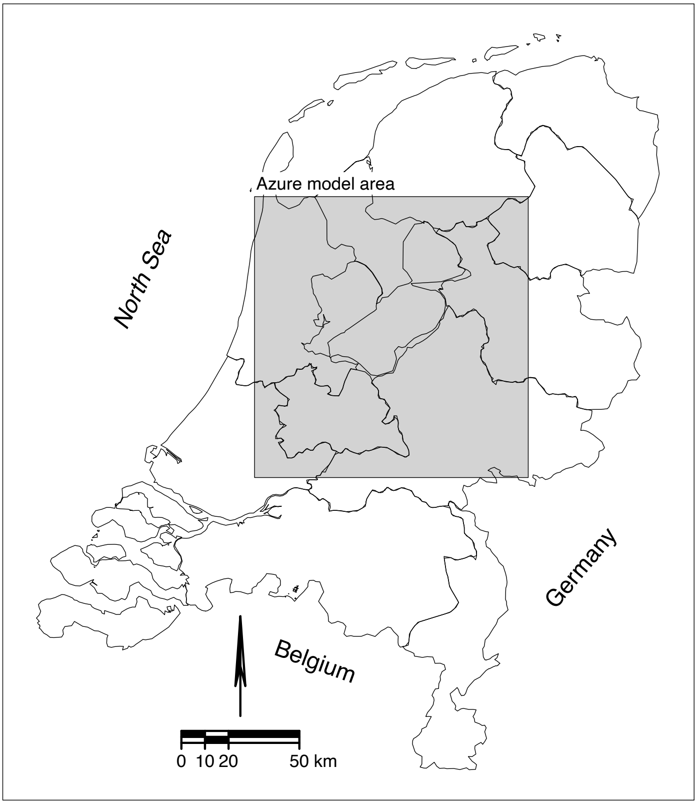

Figure 1.2: *Extent of the model area of the AZURE gorundwater flow model.*

will be for newly developed studies in the same area.

1.2 Groundwater flow model used in this thesis

In this thesis, the AZURE groundwater flow model [*de Lange and Borren*, 2014] is used. This model covers the area around and including the lake IJsselmeer, an area in the middle of the Netherlands. The model area of AZURE is shown in Figure 1.2. The subsurface data of this model is obtained from REGIS-II and GeoTOP. The GeoTOP model only incorporates the upper 50 meters of the subsurface. Therefore, only this part of the derived hydrogeological model of the AZURE model can be obtained from GeoTOP, and the deeper parts are derived from REGIS. To avoid the usage of a mixture of hydrogeological models in the to be developed feedback procedure, the calculations are applied to the deeper layers (aquifers and aquitards) of the model only, which are taken from REGIS-II.

The AZURE model is calibrated against additional information, mostly observed groundwater heads. Herewith, the derived hydrogeological model is adapted to make the model results more in agreement with observations. An important assumption is that the calibration of the groundwater flow model improves the parameterization of the derived hydrogeological model. This is a reasonable and common assumption, which makes calibrated groundwater flow models in general a source of information to improve the REGIS parameterization.

The objective of this thesis is the development of a feedback procedure making use the available information of calibrated groundwater flow models. Hence, the calibration of a groundwater flow model itself is beyond the objectives of this study.

1.3 The world ain't Gaussian, nor piecewise linear

Every observation of whatever quantity is subject to a certain degree of uncertainty. Every result derived from these observations is therefore uncertain too. Depending on the application, the uncertainty may decrease or increase, but the results are still uncertain. If the remaining uncertainty has no effect on a decision, then the uncertainty can be neglected. In many cases however, like the models at hand in this thesis, observations or derived quantities do have uncertainties which can not be ignored. An appropriate and common way to describe the uncertainty of a variable is assuming it to be random and describe the degree of uncertainty with a probability distribution [*Papoulis and Pillai*, 2002, p. 75]. The added problem is that it is often also uncertain what the exact magnitude or form of the uncertainty is. In other words, it is, for example, unknown what the variance or the shape of the probability distribution should be.

The quantification of the uncertainty of a quantity by a certain probability distribution is a choice, and therewith is the applied probability distribution a model of the uncertainty. Often, the nature of the data gives some information about the type of distribution to be used. For instance, when taking the mean value of a large number of observations, according to the central limit theorem, the probability distributions of the mean value tends to a normal or Gaussian distribution. Or if an observation is subject to a round-off error, a uniform distribution would be a safe choice to describe this error. If it is less clear which distribution should be chosen, it is also possible to fit several distributions of standard distribution families (like (log)normal, exponential, gamma or many other distributions) to the data and choose the distribution with the best fit. Examples of fit measures are the maximum likelihood fit or the Kullback-Leibler divergence. The result of such a fit is always a parametric distribution of some standard family.

Working with standard distributions is advantageous, as it often decreases the calculation effort needed. Many operations on these distributions are available in closed form solutions (analytical solutions). For instance, the maximum likelihood estimate of the parameters of a log-normal distribution can easily be obtained by the mean and standard deviation of the log-values of the observations. Or a joint distribution of Gaussian distributed random variables which is completely defined by the mean values of the marginal distributions and the covariance matrix. These properties make it very attractive to reside to standard distributions when coping with uncertainty, which is of course defendable to a certain degree.

Due to its properties, the Gaussian distribution is definitely favorite. But choosing a distribution function which diverges too much from the real function, if even known, might lead to unacceptable errors and the wrong conclusions based on the assumed uncertainties. When multiple random variables are involved in a calculation or model, the calculations may not be available in closed form solutions. And if an analytical solution could exist, it may be a tedious job to find all these solutions, and implement them, for large models.

To circumvent these problems, I decided to perform all calculations with random variables by describing the probability density functions with piecewise linear functions. Of course, a piecewise linear function is almost always an approximation of the real distribution, but often more appropriate than choosing a wrong parameterized distribution instead. Thereby, calculations are always the same for a certain operation and do not longer depend on the type of distribution. If the replacement of an analytically described distribution by a piecewise linear function is an approximation, still the operations on these functions are analytical. This means, the operations are performed on linear functions within a certain interval (bin). These operations have analytical solutions within a bin or between bins of functions of different variables. Therefore, the method of piecewise linear functions can be considered as a hybrid numerical-analytical method.

1.4 Research objectives and thesis outline

As mentioned above, the general purpose model REGIS provides subsurface information to groundwater flow models. This abstracted information, the derived hydrogeological model, contains usually aggregated data from the REGIS model. This derived hydrogeological model is often improved by calibrating the groundwater flow model against other available data. Such a calibrated derived hydrogeological model contains valuable information from which the REGIS model could benefit. Often, one layer (aquifer or aquitard) of the derived hydrogeological model consists of n units of the REGIS model, then $2n$ parameters (layer thickness and conductivity of each REGIS unit) are involved in the aggregated parameter of the derived hydrogeological model (transmissivity for aquifers or vertical hydraulic resistance for aquitards). Due to the aggregation of layers no simple deterministic feed back procedure is possible, and currently, no formal or objective method is available to lead these improvements back into REGIS. Therefore, the main objective of this study is:

> *Develop a method or procedure to let the generic hydrogeological model, in our case REGIS, benefit from the improvements of a calibrated groundwater flow model.*

This is a broad stated aim and needs to be broken into sub-objectives.

If a deterministic procedure is beyond reach, a stochastic feed back method has to be considered. It is recognized that uncertainty is ubiquitous in all models and data, and this uncertainty is usually described and quantified by probability distributions. However, these distributions of the uncertain data, the random variables, usually belong to a variety of families of standard distributions, or bear any nonstandard distribution. Often, operations on these random variables do not yield any closed form (analytical) solution. For instance, in the current study the transmissivity is formed by the product of the conductivity (k) and the layer thickness (D). The conductivity is usually assumed to be log-normal distributed but the layer thickness not. Their product has thus presumably no standard form. Therefore, the feedback procedure must account for uncertainty, preferably for all kinds of distributions. So a sub-objective is:

> *Develop a method which accounts for uncertain data of all kinds of probability distributions.*

REGIS is a multipurpose hydrogeological model, which holds that in the same area multiple groundwater flow models may be available. Since every calibrated groundwater flow model may contribute to the improvement of the subsurface data, the method must be able to handle this. Thereby, groundwater flow models can be developed with a different objective, so the uncertainty of the calibrated result may differ. So another sub-objective is:

> *Develop a method which can use multiple calibrated groundwater flow models in the same area and with different uncertainty.*

Figure 1.3 provides the thesis outline in terms of different chapters and how they fit in the general aim of improving a general purpose hydrogeological model by the feedback of calibration results from regional or local groundwater models.

In Chapter 2 a method is developed to perform calculations with piecewise linear probability density functions, which is used in the rest of this thesis, to account for uncertainty in data and any derived properties from these data. As a first proof of application it is applied to the problem of aggregation and upscaling of conductivities to transmissivities, which is common in deriving hydrogeological models for local groundwater models. Here, layers and conductivities in boreholes are first aggregated to local transmissivities at borehole locations and then interpolated with ordinary kriging. Using piecewise linear approximations, the probability distributions of transmissivity at the interpolation locations are directly calculated.

In Chapter 3 a method is presented to find the most likely values of an uncertain quantity, given observations. The observations are the calibrated vertical resistances of an aquitard, and the quantities are the layer thicknesses and conductivities of sub-layers which build-up this aquitard. This procedure is applied to two distinct areas in Chapter 4 to test if the method is able to find lateral differences in the updated conductivity values, which were initially assumed to be spatially uniform.

The method in Chapter 3 is able to find one most likely parameter value, given one observation (one value of the calibrated resistivity in each grid cell). In Chapter 5 the problem is redefined in a Bayesian context, which allows for estimating the full posterior probabilities. Thus the updated probability distributions of layer

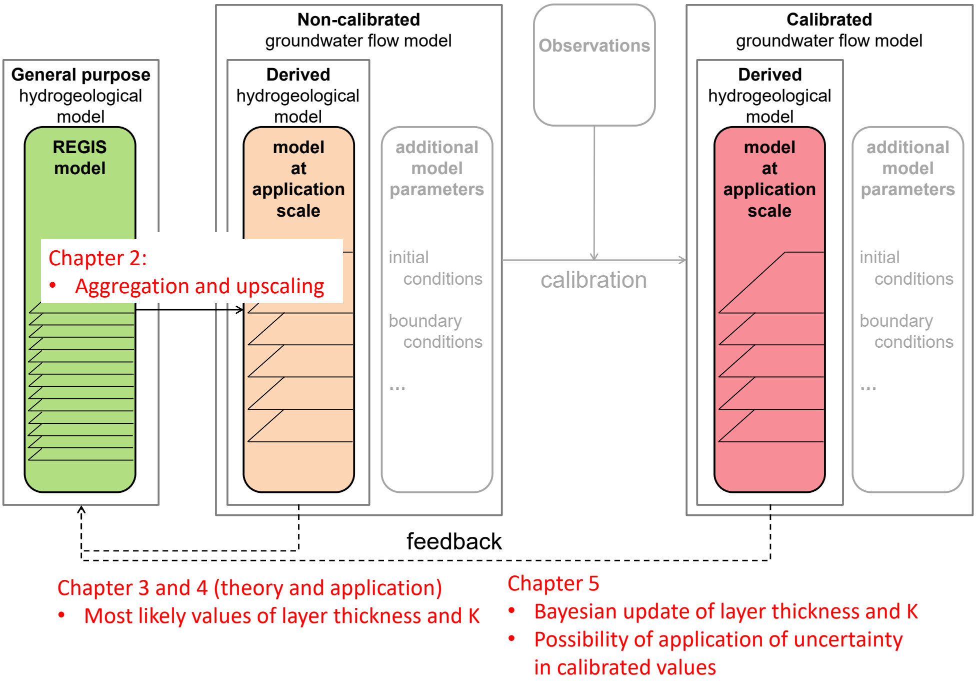

Figure 1.3: *Thesis outline in terms of the general problem definition.*

thicknesses and conductivities of the hydrogeological model are estimated, given the calibrated resistivity. An added advantage of this approach is that multiple observations can be used at the same location, and that observations can be uncertain. This means that multiple uncertain resistivity values of multiple calibrated groundwater models in one region can be used to improve the generic hydrogeological model.

Finally, in Chapter 6 the results are discussed and recommendations for future research are given.

2

Uncertainty propagation with probability density functions using piece-wise linear approximations

Abstract. In many fields of study, and certainly in hydrogeology, uncertainty propagation is a recurring subject. Usually, parameterized probability density functions (PDFs) are used to represent data uncertainty, which limits their use to particular distributions. Often, this problem is solved by Monte Carlo simulation, with the disadvantage that one needs a large number of calculations to achieve reliable results. In this paper, a method is proposed based on a piecewise linear approximation of PDFs. Herewith, the uncertainty propagation with these discretized PDFs is distribution independent. The method is applied to the upscaling and interpolation of conductivity data, and carried out in two steps: the vertical upscaling of conductivity values from borehole data to aquifer scale, and the spatial interpolation of the transmissivities. The results of this first step are complete PDFs of the transmissivities at borehole locations reflecting the uncertainties of the conductivities and the layer thicknesses. The second step results in a spatially distributed transmissivity field with a complete PDF at every grid cell. We argue that the proposed method is applicable to a wide range of uncertainty propagation problems.

This chapter is adapted from Lourens, A., and F.C. van Geer, Uncertainty Propagation of Arbitrary Probability Density Functions Applied to Upscaling of Transmissivities, Stochastic Environmental Research and Risk Assessment, *30, 237–249, 2016*

SUBSURFACE PARAMETERS are essential data for groundwater flow models. Often, these data originate from borehole descriptions in which thin layers (core scale) are distinguished based on lithological and sedimentological information. The thickness of these layers may vary from a few centimeters up to several meters, depending on the subsurface structure and the drilling method. Typically, the described layers are vertically aggregated to aquifer and aquitard classes at a scale which fits the groundwater model requirements. This scale will be referred to as point scale. The thickness of aquifers typically comes on the order of a few meters to 100 m or up. The core scale layers are normally populated with hydraulic conductivities derived from the literature or estimated in the laboratory. Next, point values of transmissivities and resistances are calculated by vertical integration of the conductivity values. Subsequently, these point values are interpolated to acquire a spatial parameter field at model scale. This scale has a lateral block size of about 100 m to 1,000 m.

An important issue in the upscaling procedures is the uncertainty of the model parameters. This uncertainty can be divided into two sources. Firstly, the available observations, at core scale, are uncertain, introducing uncertainty in the upscaling to point scale values. In this case, each observation is not treated as one known value but as a random variable (RV). Secondly, there is uncertainty about the spatial distribution of the parameter. At observed locations the point scale parameter values are the upscaled RVs. At unobserved locations, assumptions have to be made about the spatial structure. This spatial structure can be described by regionalized variables (ReV) [*Journel and Huijbregts*, 1978, p. 26].

In the Netherlands, a large database (REGIS) exists [*Vernes et al.*, 2005; *Vernes and van Doorn*, 2006], in which all differentiated layers from all boreholes are described at core scale by litho-stratigraphical units. Ranges of possible parameter values for hydraulic conductivity and porosity are assigned to these units. For REGIS, these ranges are obtained from laboratory tests and literature search. When a sufficient amount of data is available for a litho-stratigraphical unit, a probability distribution is derived for the parameter of this unit. In this article, these probability distributions are used to measure the uncertainty about the hydraulic conductivities at core scale.

As described extensively in the literature, the upscaling of hydraulic parameters is far from trivial and depends highly on: the support scale of the observations, the required model scale, the presence of anisotropy in the hydraulic conductivity, and boundary conditions of the flow problem at hand [*Dagan*, 1986; *Bierkens and Weerts*, 1994; *Tran*, 1996; *Fiori et al.*, 2011]. Some clear overviews about these subjects are given by *Cushman et al.* [2002]; *Nœtinger et al.* [2005]; *Sanchez-Vila et al.* [2006]. Upscaling of hydraulic conductivities needs different approaches in one, two and three dimensions. With an increasing number of dimensions the complexity of the upscaling method increases even more. The upscaled one-dimensional conductivity is calculated by the harmonic mean. In isotropic media with a two-dimensional schematization, the upscaled conductivity can be obtained by the geometric mean [*De Wit*, 1995; *Hristopulos*, 2003]. The three-dimensional upscaling is much more complicated and many upscaling methods are proposed in the literature [*King*, 1989; *De Wit*, 1995; *Hristopulos and Christakos*, 1999; *Hristopulos*, 2003; *Boschan and Nœtinger*, 2012]. Although in two dimensions the geometric mean yields a usable effective conductivity in isotropic media, in strong heterogeneous media the result may divert too much from realistic values. For the latter case, different solutions are proposed in the literature for strong heterogeneous or binary media [*King*, 1989; *Pancaldi et al.*, 2007; *Boschan and Nœtinger*, 2012]. Block kriging on log-conductivity values is equal to geometric upscaling of the two-dimensional situation. If the correlation length is larger then the block size, the within block variability will be low. In this case, block kriging will yield accurate effective conductivity values. Subsequently, these block average values, the model scale, can be used as a starting point in the above mentioned upscaling methods. In the upscaling literature, this scale is often denoted as the fine scale grid.

In this article, the vertical one-dimensional upscaling is used at point scale, and the lateral two-dimensional upscaling is applied using kriging interpolation. In both cases, the complete parameter distributions of the observation data, as stored in the REGIS database, are used. Herewith, the probability density functions at each grid cell are calculated. These parameter distributions are assumed to be representative at the model scale.

This article is not meant as a contribution to the problem of scale dependent hydraulic conductivities but as a description of a method to propagate uncertainties. Nevertheless, the proposed method can be used in conjunction with the above mentioned upscaling methods, thus propagating the observation uncertainty, but this is left for future work.

In this article, we will focus on the upscaling of hydraulic conductivities to transmissivities. To be useful to groundwater models, the point scale conductivities, which in fact are RVs, have to be upscaled to spatial distributed transmissivities. Commonly, only one measure of this RV (e.g., mean) is used to perform this upscaling. Herewith, only information about the uncertainty of the interpolated mean is obtained, disregarding the uncertainty of the observations. Techniques like Monte Carlo simulation (MC) are often used to obtain results reflecting the data uncertainty. However, a disadvantage of MC is the dependence of the number of calculations, the sampling strategies used [*Kyriakidis and Gaganis*, 2013], and the large number of calculations needed to obtain reasonable results.

The objective of our study is twofold: the derivation of a method to perform uncertainty propagation calculations with complete PDFs, and the application of this method in the upscaling and spatial interpolation of subsurface parameters. To take full advantage of the prior knowledge of the uncertainty of data, we present a method to propagate this uncertainty throughout all the calculations. Since the RVs are not described by their statistical moments but by numerically discretized PDFs, the proposed method is applicable regardless of the type of distributions used. Although the described technique can be used in conjunction with techniques that account for anisotropy, the proposed methods are applied to homogeneous examples.

The developed method is described in Section 2.1. In Section 2.2 the method is applied to the upscaling of real world borehole data to transmissivities at model scale, using kriging interpolation. The performance of the method is compared with an MC calculation. Section 2.3 contains the discussion and conclusions.

2.1 Methodology

Parameters obtained from observations are always subject to uncertainty. When this uncertainty contributes significantly to the result of calculations, it should be accounted for. A generally applicable method to propagate the uncertainty of random variables (RV) in a wide range of calculations is very attractive. This method should be independent of the shape of probability density functions (PDF) and supports binary operations $(+ - \times /)$ and elementary functions. In this section, we first develop a method to perform calculations with discretized PDFs. Thereafter, this method is implemented in the vertical upscaling of core scale conductivities. Finally, the method is integrated in the kriging interpolation to obtain the PDF of the spatial distributed transmissivity data reflecting all sources of uncertainty.

2.1.1 Piecewise linear PDFs

Commonly, parameterized PDFs are used to perform uncertainty calculations analytically. This means that for every possible combination of types of PDFs an analytical solution must be available. When many types of PDFs and operations need to be supported, numerous derivations have to be made. For long chains of calculations, this is highly inefficient. Moreover, the resulting PDFs should be known in closed analytical form, which can not always be achieved [*Holmes and Buhr*, 2007; *Silverman et al.*, 2004a].

We aim at a method which is universally applicable and independent of the type of distribution used. To achieve this, a combination of a numerical and an analytical approach is used, that is, the PDFs are described numerically and the arithmetic is performed analytically. A common way to discretize PDFs is to describe them piecewise linear [*Kaczynski et al.*, 2012; *Vander Wielen and Vander Wielen*, 2015]. Herewith,

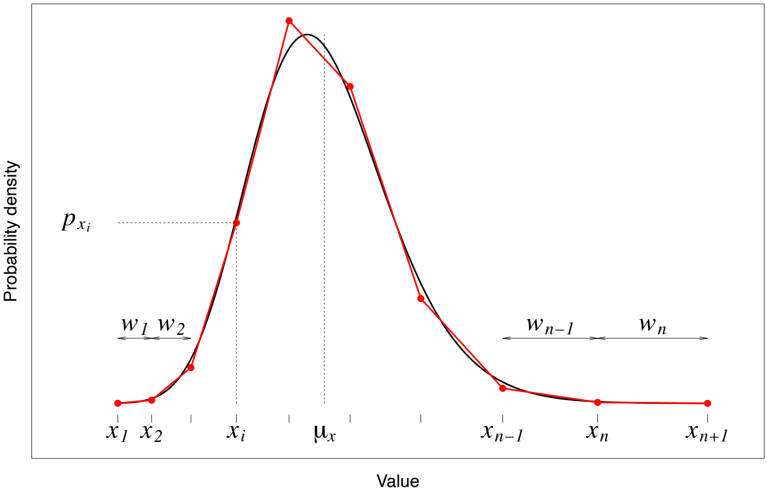

Figure 2.1: Example of a piecewise linear discretization of a PDF. The discretized PDF (red) is a n bins discretization of the real PDF (black). At the red points, the cumulative probabilities are equal to those of the real PDF. In this picture is: x_i the value of the PDF, p_{x_i} the probability density at value x_i, w_i the width of bin i, and μ_x the average value of the PDF.

any probability distribution which can be approximated by a piecewise linear PDF can be used. A drawback of this method is the introduction of inaccuracies by linearization, and the need for truncation of distributions with a one or two sided infinite domain. However, this drawback can largely be overcome by the choice of a sufficient number of discretization points, and discretize large tails when needed. In Figure 2.1 an example of a piecewise linear PDF is given. Between two discretization points, the PDF is described by a linear function. This interval is referred to as a bin [*Izenman*, 1991]. A calculation method with discretized PDFs is described before in *Jaroszewicz and Korzeń* [2012] and *Korzeń and Jaroszewicz* [2014]. However, their approach is different from ours which makes both methods applicable in different types of problems. A comparison of both methods is described in Section 2.2.2.

2.1.2 Calculations with PDFs

Binary operations

When the PDF of an RV can be described analytically, the result of a binary operation $(+ - \times /)$ can be described analytically as well. Let Z be the RV formed by the joint distribution of two independent RVs X and Y. The general formulation of the cumulative distribution function (CDF) of Z can be described as *Papoulis* [1991, p. 132 ff]

$$F_z(z) = \int\int f_x(x) f_y(y)\,\mathrm{d}x\,\mathrm{d}y, \tag{2.1}$$

where $f_x(\cdot)$ and $f_y(\cdot)$ are the PDFs of X and Y, respectively. In this equation, the integration boundaries depend on the value of z and the binary operation to be calculated. Let Z be the sum of X and Y, then the probability $\Pr\{Z < z\}$ can be written as

$$F_z(z) = \int_{y=-\infty}^{\infty} \int_{x=-\infty}^{z-y} f_x(x) f_y(y) \,\mathrm{d}x \,\mathrm{d}y. \tag{2.2}$$

The integration boundaries for subtraction, multiplication and division are given in Appendix A. Unfortunately, for piecewise linear PDFs such analytical formulation can not be solved as one integral. However, the PDF of each bin of the PDFs can be described analytically. So for each bin of the marginal distributions, the linear functions $f_{x,i}(\cdot)$ and $f_{y,j}(\cdot)$ can be defined as

$$f_{x,i}(x) = p_{x_i} + r_{x_i}(x - x_i) \qquad \text{for } x \in \langle x_i, x_{i+1}] \tag{2.3}$$

$$f_{y,j}(y) = p_{y_j} + r_{y_j}(y - y_j) \qquad \text{for } y \in \langle y_j, y_{j+1}], \tag{2.4}$$

where p_{x_i} and p_{y_j} are the probability densities at the values x_i and y_j, respectively. The slopes of these functions are defined as $r_{x_i} = (p_{x_{i+1}} - p_{x_i})/(x_{i+1} - x_i)$ and $r_{y_j} = (p_{y_{j+1}} - p_{y_j})/(y_{j+1} - y_j)$. With these functions, we can define the piecewise analytical solution of the CDF of Z by integration of the probability density of the area inside the joint bin below the line $z = x + y$. The integration area is split up into four sub-areas as can be seen in Figure 2.2. Because X and Y are independent, the probability of the rectangle sub-area a can be easily defined by the product of its marginal probabilities

$$F_{z,ij,a}(z) = \Pr\{x_i < X \leq x_{l,i}\} \Pr\{y_j < Y \leq y_{l,j}\}. \tag{2.5}$$

Equivalently, the probabilities of area b and c are expressed. The equation of the probability of sub-area d of joint bin (i, j) can be written as

$$F_{z,ij,d}(z) = \int_{y=y_{l,j}}^{y_{u,j}} \int_{x=x_{l,i}}^{z-y} f_{x,i}(x) f_{y,j}(y) \,\mathrm{d}x \,\mathrm{d}y. \tag{2.6}$$

The integration boundaries $y_{l,j}$, $y_{u,j}$, $x_{l,i}$ and $z - y$ are portrayed in Figure 2.2. When $z > y_{j+1} + x_{i+1}$ or $z < y_j + x_i$, the line $z = x + y$ does not intersects the joint bin (i, j). Therefore, z_{ij} is defined to replace z in the calculations of joint bin (i, j). The value of z_{ij} is calculated using $z_{ij} = \min(\max(z, x_i + y_j), x_{i+1} + y_{j+1})$. Integration of Equation (2.6) yields (see Appendix A.1.1 for its derivation)

$$\begin{aligned} F_{z,ij,d}(z_{ij}) = \; & \tfrac{1}{2} p_{x_{l,i}} p_{y_{u,j}} (y_{u,j} - y_{l,j})^2 - \tfrac{1}{3} p_{x_{l,i}} r_{y_j} (y_{u,j} - y_{l,j})^3 \\ & + \tfrac{1}{6} r_{x_i} p_{y_{u,j}} (y_{u,j} - y_{l,j})^3 - \tfrac{1}{8} r_{x_i} r_{y_j} (y_{u,j} - y_{l,j})^4. \end{aligned} \tag{2.7}$$

To obtain the cumulative probability for a particular value of Z, a summation of the probabilities of all joint bins is performed

$$F_z(z) = \sum_{j=1}^{n_y} \sum_{i=1}^{n_x} \sum_{A=a,b,c,d} F_{z,ij,A}(z), \tag{2.8}$$

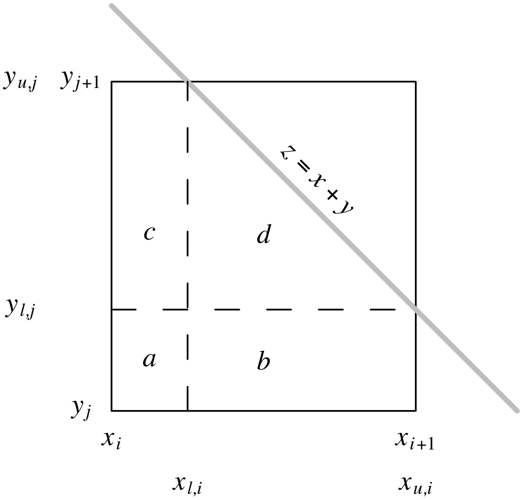

Figure 2.2: Integration boundaries of the piecewise analytical CDF. Shown is the dependence of the integration boundaries on the position of the line z in the box of the joint bin (i, j).

where n_x and n_y are the numbers of bins of X and Y, respectively.

From Equation (2.7) the PDF of Z can be derived by taking the first derivative with respect to z. The parameters depending on z have to be rewritten as a function of z as $x_{u,i} = z - y_{l,j}$, $y_{u,j} = z - x_{l,i}$ and $p_{y_{u,j}} = f_{y,j}(z - x_{l,i})$. Herewith the derivative yields

$$\begin{aligned} f_{z,ij,d}(z) = \; & p_{x_{l,i}} p_{y_{u,j}} (y_{u,j} - y_{l,j}) - \tfrac{1}{2} p_{x_{l,i}} r_{y_j} (y_{u,j} - y_{l,j})^2 \\ & + \tfrac{1}{2} r_{x_i} p_{y_{u,j}} (y_{u,j} - y_{l,j})^2 - \tfrac{1}{3} r_{x_i} r_{y_j} (y_{u,j} - y_{l,j})^3 . \end{aligned} \tag{2.9}$$

The PDF of all bins writes

$$f_z(z) = \sum_{j=1}^{n_y} \sum_{i=1}^{n_x} f_{z,ij,d}(z). \tag{2.10}$$

Analogous to the summation, the integration can also be performed for subtraction, multiplication and division. An illustration of the equi Z-lines of four binary operations is given in Figure 2.3. The derivations of the four binary operations can be found in Appendix A.

Discretizing unknown variable Z

Performing a binary operation like Equation (2.8), raises the need for a proper discretization of the unknown RV Z. Due to linearization, the integral of this PDF will usually not describe the CDF exactly. This probability error for each bin has to be as small as possible without increasing the number of bins too much.

An algorithm is proposed which starts with at least three predefined Z-values (e.g., z_{min}, z_{max}, and z_{mean}). Subsequently, new Z-values are added during calculation. For every Z-value, the cumulative probability (Equation (2.8)) and the

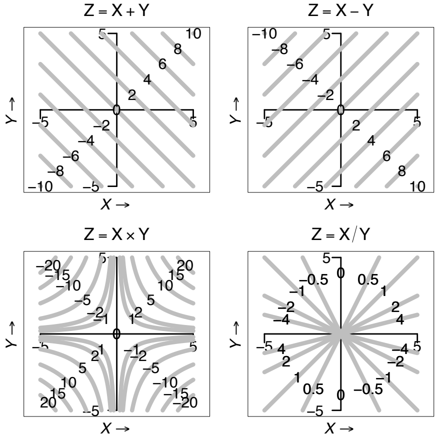

Figure 2.3: Example of the graphical representation of CDFs *of four binary operations between two independent* RVs. *The gray lines are the upper boundaries of the integration area of the cumulative probability for a certain value of Z.*

probability density (Equation (2.10)) are calculated. The probability of each bin can now be calculated in two ways: the difference of the cumulative probability at each edge of the bin, and the integration of the linearized probability density of the bin. Herein, the first probability is the exact solution of the calculations and the second method yields an approximate value. The difference between these probabilities is the error caused by the linearization of the PDF. The bin with the largest absolute probability error will be split up at its center of mass of the probability of the linearized function. This algorithm runs until all probability errors are smaller then a certain threshold, or a predefined maximum number of bins is reached. In Figure 2.4, an example of one iteration of the summation of two independent RVs (both $\mathcal{N}(2,1)$) is illustrated.

2.1.3 Construction of probability fields of transmissivity

This section describes a two step approach of the construction of probability fields of transmissivity. Firstly, the borehole data is upscaled to aquifer scale at point locations. Secondly, these upscaled values are horizontally interpolated using kriging interpolation. Both steps make use of the calculation methods as described in Section 2.1.2.

Vertical upscaling

The transmissivity of a layer at core scale is calculated from borehole data by multiplying the layer thickness by the conductivity

$$T_l = K_l(L_l - L_{l+1}), \tag{2.11}$$

where index l denotes the layer number, T_l is the transmissivity and K_l the hydraulic conductivity of layer l, and L_l the height of the top of layer l, measured relative to, for example, Amsterdam Ordnance Datum. The layer numbers increase

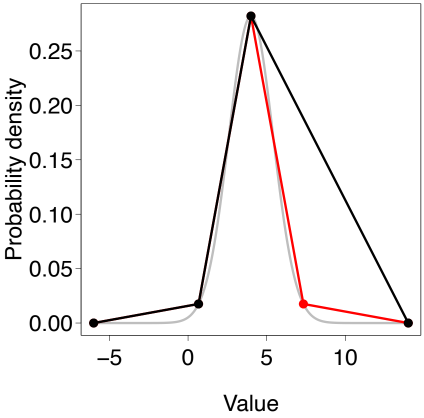

Figure 2.4: Refining the PDF *by adding a Z-value. The gray line is the true solution, the black line shows the 4-point* PDF*, and the red line shows the effect of adding the* 5^{th} *defined Z-value.*

downwards, so the bottom of layer l coincides with the top of layer $l+1$ (i.e., L_{l+1}). Subsequently, the upscaled aquifer transmissivity at point scale is defined by

$$T = \sum_{l=1}^{n} T_l, \tag{2.12}$$

where n is the number of layers, at core scale, which are combined to one aquifer. Equation (2.12) only holds for horizontal flow within an aquifer. As denoted in the introduction of this chapter, we assume the conductivity parameter values appropriate for the scale used after upscaling. Subjects like anisotropy are beyond the scope of this article.

Both, the layer thickness and the hydraulic conductivity are subject to uncertainty. When transmissivities are upscaled from consecutive layers, these individual transmissivities are correlated because of the uncertainty of the boundaries between these layers. In order to perform the summation of transmissivities correctly, we need to know the correlation between the layers. The covariance of the transmissivities of two consecutive layers can be calculated as

$$\begin{aligned} \operatorname{cov}(T_l, T_{l+1}) &= \operatorname{cov}(K_l(L_l - L_{l+1}), K_{l+1}(L_{l+1} - L_{l+2})) \\ &= \quad \operatorname{cov}(K_l L_l, K_{l+1} L_{l+1}) - \operatorname{cov}(K_l L_l, K_{l+1} L_{l+2}) \\ &\quad - \operatorname{cov}(K_l L_{l+1}, K_{l+1} L_{l+1}) + \operatorname{cov}(K_l L_{l+1}, K_{l+1} L_{l+2}). \end{aligned} \tag{2.13}$$

When we assume all variables K and L mutually independent, only the third covariance ($-\operatorname{cov}(K_l L_{l+1}, K_{l+1} L_{l+1})$) is not equal to 0. According to *Bohrnstedt and Goldberger* [1969] this covariance can be written as

$$\operatorname{cov}(K_l L_{l+1}, K_{l+1} L_{l+1}) = \mathrm{E}[K_l]\,\mathrm{E}[K_{l+1}]\operatorname{var}(L_{l+1}). \tag{2.14}$$

The correlation coefficient can now be written as

$$\rho_{(T_l,T_{l+1})} = -\frac{\mathrm{E}[K_l]\,\mathrm{E}[K_{l+1}]\,\mathrm{var}(L_{l+1})}{\sqrt{\mathrm{var}(T_l)\,\mathrm{var}(T_{l+1})}}. \tag{2.15}$$

If the value of $\rho_{(T_l,T_{l+1})}$ can not be neglected, we have to account for correlations in Equation (2.12). When the correlations differ significantly from 0, also in the calculations of Section 2.1.2 the correlations should be taken into account. The correlations as calculated from the observation data are found in Section 2.2.1.

Horizontal upscaling: semivariogram

Sample semivariograms are usually derived from observations which are assumed to be deterministic values. Since our point scale observations are RVs, this will cause a different sample semivariogram and the way it is obtained. Our aim is to find a semivariogram based on uncertain observations and to find the PDF of the interpolation. Although the observations are of a different nature then usual (RVs instead of deterministic), we assume the intrinsic hypothesis [*Journel and Huijbregts*, 1978, p. 11] still holds.

The definition of the semivariogram is [*Goovaerts*, 1997, p. 96]

$$\gamma(h) = \tfrac{1}{2}\,\mathrm{E}[(Z(u) - Z(u+h))^2], \tag{2.16}$$

where $Z(u)$ is the sample value at location u, and h is the spacing between two observation locations. Equation (2.16) can be rewritten as

$$\gamma(h) = \mathrm{E}\left[\left(\tfrac{1}{\sqrt{2}}(Z(u) - Z(u+h))\right)^2\right] = \mathrm{E}\left[\Delta_Z(h)^2\right]. \tag{2.17}$$

From the intrinsic hypothesis it follows that $\Delta_Z(h)$ has a symmetrical distribution function with zero mean. So $\Delta_Z(h)$ is the RV with a probability distribution describing the difference between two observations at lag h, scaled with factor $1/\sqrt{2}$. Equation (2.17) can now be written as $\gamma(h) = \mathrm{var}(\Delta_Z(h))$. The PDF of $\Delta_Z(h)$ is derived from the observations $Z(u)$, which can be either deterministic values or RVs. The effect of the observations being RVs is shown in Figure 2.5. As expected, a nugget effect arises from the use of RVs as observations.

In general, $\Delta_Z(h)$ is assumed to be Gaussian distributed, which is not always the case [*Journel and Huijbregts*, 1978, p. 50]. In the procedure described here, the shape of the distribution is derived from the observations. The assumption we make is that the shape of $\Delta_Z(h)$ is independent of h, only the variances differ.

Since we want to use the distribution of $\Delta_Z(h)$ in the kriging interpolation, we have to relate it to the covariance function. For a stationary random function, the covariance function and the correlogram are directly related to the semivariogram [*Journel and Huijbregts*, 1978, p. 32]. The covariance function can be written as

$$C(h) = C(0) - \gamma(h), \tag{2.18}$$

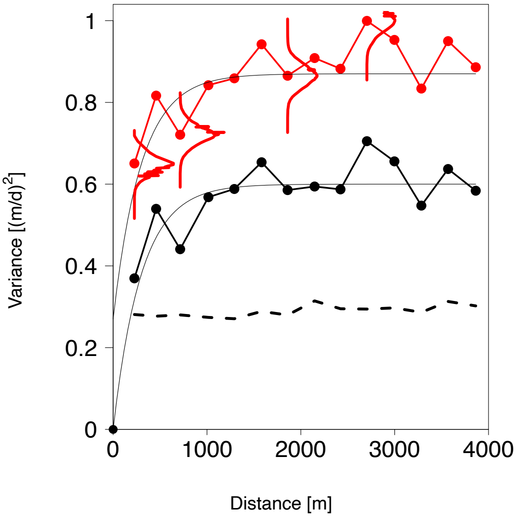

Figure 2.5: *Example of a sample semivariogram. The black lines show the result when the observations are treated as deterministic values. The red line is the result of observations treated as* RV*s. The dashed line shows the difference between the red and the black line, which is the expected nugget effect. The smooth black lines are the fitted variogram models. At four points the* PDF *of $\Delta_Z(h)$ is drawn from which the variance is derived. The semivariogram is derived from the log-values of the observations.*

where $C(h)$ is the covariance at lag h, with $C(0) = \gamma(h \to \infty) = \mathrm{var}(\Delta_Z(h \to \infty))$. For convenience we define $\Delta_Z = \Delta_Z(h \to \infty)$. The correlogram is defined as

$$\rho(h) = \frac{C(h)}{C(0)}, \tag{2.19}$$

where $\rho(h)$ is the correlation coefficient at lag h. From Equation (2.19) we can write

$$C(h) = \rho(h)C(0) = \rho(h)\,\mathrm{var}(\Delta_Z). \tag{2.20}$$

From this relation we derive that the covariance $C(h)$ can be calculated as

$$C(h) = \mathrm{var}\left(\sqrt{\rho(h)}\Delta_Z\right). \tag{2.21}$$

The covariance functions must be positive definite [*Journel and Huijbregts*, 1978, p. 34], so $\rho(h) \geq 0$.

Horizontal upscaling: interpolation

The vertical upscaled borehole data, as described in Section 2.1.3, are used in spatial interpolation. Since these data are subject to uncertainty, an interpolation technique which can handle this kind of data must be chosen. We applied ordinary kriging to perform this interpolation. In this section we describe the way we incorporate the uncertainty of the observations, including the shape of the distributions, in the kriging variance.

Ordinary kriging is based on two equations [*Isaaks and Srivastava*, 1989, p. 280 ff]. The interpolation of the observation values is described by

$$\hat{Z}(u_0) = \sum_{\alpha=1}^{n} \lambda_\alpha Z(u_\alpha), \tag{2.22}$$

where $\hat{Z}(u_0)$ is the kriging estimate at the unsampled location u_0, λ_α the weight factor of $Z(u_\alpha)$, and n the number of sample locations used in the estimate. The variance of $\hat{Z}(u_0)$ is described by

$$\operatorname{var}(\hat{Z}(u_0)) = \sum_{\alpha=1}^{n} \sum_{\beta=1}^{n} \lambda_\alpha \lambda_\beta C(h_{\alpha\beta}), \tag{2.23}$$

where $C(\cdot)$ is the covariance function as discussed in Section 2.1.3, and $h_{\alpha\beta}$ is the distance between location u_α and u_β.

In general, $Z(u_\alpha)$ represents a deterministic value at each location, which yields a deterministic value $\hat{Z}(u_0)$ as well. The variance of $\hat{Z}(u_0)$ is calculated by Equation (2.23), and if probabilities are calculated $\hat{Z}(u_0)$ is assumed to have a Gaussian distribution. Together, these two results describe the conditional PDF at the interpolation location (conditional to the values found at the observation locations).

Since we have PDFs available at all sample locations we use these PDFs in Equation (2.22). This yields an RV for $\hat{Z}(u_0)$ which honors the uncertainty, including the distribution, of the sample data. Additionally, we want to use the distribution of Δ_Z in the uncertainty of the interpolation. In Section 2.1.3 we presented a method to obtain the PDF of $C(\cdot)$, described in Equation (2.21). Inserting Equation (2.21) in Equation (2.23) yields

$$\begin{aligned} \operatorname{var}(\hat{Z}(u_0)) &= \sum_{\alpha=1}^{n} \sum_{\beta=1}^{n} \lambda_\alpha \lambda_\beta \operatorname{var}\left(\sqrt{\rho(h_{\alpha\beta})}\Delta_Z\right) \\ &= \operatorname{var}\left(\sum_{\alpha=1}^{n} \sum_{\beta=1}^{n} \sqrt{\lambda_\alpha \lambda_\beta \rho(h_{\alpha\beta})}\Delta_Z\right). \end{aligned} \tag{2.24}$$

Herein, $\sum_{\alpha=1}^{n} \sum_{\beta=1}^{n} \sqrt{\lambda_\alpha \lambda_\beta \rho(h_{\alpha\beta})}\Delta_Z$ is the RV describing the uncertainty of the interpolation with a distribution based on Δ_Z. When added to $\hat{Z}(u_0)$, the resulting RV describes the probability distribution of the interpolation.

2.2 Results

2.2.1 Application to real world data

This section shows an example of upscaling and interpolation of borehole data, using the proposed methods. From the REGIS database of the Geological Survey of the Netherlands, we used data from the Kiezeloöliet Formation from an area in the south of the Netherlands. The dataset contains about 200 boreholes with data

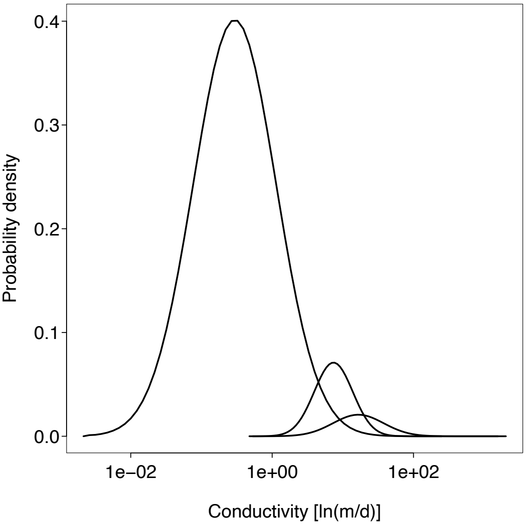

Figure 2.6: *PDFs of three classes of sand as used with the upscaling of the borehole data. From left to right: fine sand, medium fine sand, and coarse sand. The horizontal axis is logarithmic which explains the apparent difference in integrated area.*

from the second aquifer [*Vernes et al.*, 2005]. This aquifer consists mainly of sandy deposits which are divided into three classes with significant different conductivity distributions. Figure 2.6 shows the PDFs of these distributions.

The vertical upscaling of the borehole data is performed as described in Section 2.1.3. The number of core scale layers at one borehole varied between 1 and 40 layers with an average of about 9 layers. During upscaling, we calculated 1645 correlations between consecutive layers using Equation (2.15). It appears that almost all (1638) correlations between the transmissivities of consecutive layers have a value between -0.05 and 0, the rest has values between -0.085 and -0.05. Because of these low correlations, we performed the upscaling without taking the correlations into account.

The variogram model, as shown in Figure 2.5, is derived from the upscaled borehole data. The PDFs of the conductivities are log-transformed before kriging [*Journel and Huijbregts*, 1978, p. 570] and the interpolated PDFs are back transformed afterwards. In this example we used an exponential variogram with range 300 m, sill 0.6 $\ln(\mathrm{m/d})^2$, and nugget 0.27 $\ln(\mathrm{m/d})^2$.

The performance of the PDF calculation used at interpolation of uncertain data, by using Equation (2.22), is compared to a Monte Carlo (MC) simulation. For this purpose, we draw a large number of random realizations (n_{MC}) of the PDFs of the observations. These random realizations are treated as observations in kriging. Since we assume that the semivariogram does not alter for each realization, the same sets of weight factors, λ_α, are used for both, the PDF and the MC calculations. Subsequently, the results of MC are transformed to a CDF and PDF, as displayed

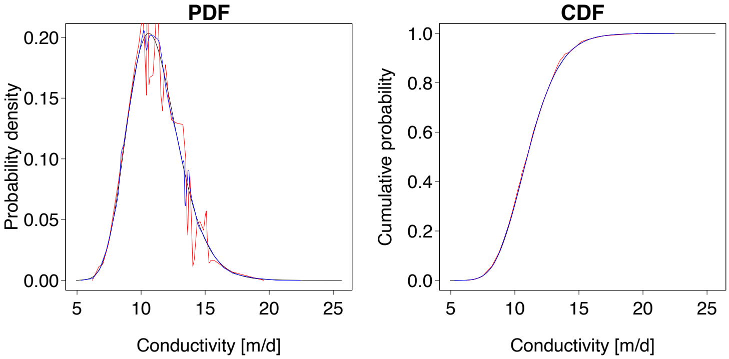

Figure 2.7: *Result of PDF calculations compared to MC. Black: PDF calculations, red: MC with* $n_{MC} = 1,000$*, blue MC with* $n_{MC} = 20,000$*. The black and blue line coincide.*

in Figure 2.7. It can be seen that the CDFs of both MC runs ($n_{MC} = 1,000$ and $n_{MC} = 20,000$) fit quite well with the CDF of the PDF calculations. However, the PDFs of the MC calculations are less smooth than the PDF of the PDF calculations. The interpolated location in this example is the same location as in Figure 2.8 indicated by a red circle.

Some results of the kriging interpolation are shown in Figure 2.8. The results in this example are obtained by point kriging. At every kriging location, two PDFs are drawn. The dashed line PDFs are the results of kriging applied on deterministic observations, assuming the underlying stationary random function of conductivities to have a log-normal PDF. The solid lines are the kriging results with observations as random variables as described before. As expected, the widths of the PDFs of the interpolated point data are smaller than those of the interpolated PDFs. This is caused by the added uncertainty of the observations in the case of the PDF interpolations.

Another method to include uncertainty of the observations is kriging with uncertain data [*Marsily*, 1986, p. 299]. There, the error variance of each observation is subtracted from the corresponding main diagonal element of the kriging matrix. The differences between both methods is that kriging with uncertain data relies on Gaussian distributions, where the proposed method can handle distributions of any shape.

2.2.2 Comparison of calculation methods

Performing calculations with RVs of which the descriptions of the PDFs are expressed by simplified functions, is described before in the literature. In this section, the main differences between the calculation method of *Jaroszewicz and Korzeń* [2012] and the piecewise linear method, as described in this chapter, are discussed.

Both methods divide the PDFs in intervals where the probability densities are approximated by one or more polynomial functions. The piecewise linear method uses only one linear function, where the method of Jaroszewicz and Korzeń uses also higher order polynomials, implemented as Chebyshev polynomials. The latter method has the ability to describe the curve of the PDF much more accurately than

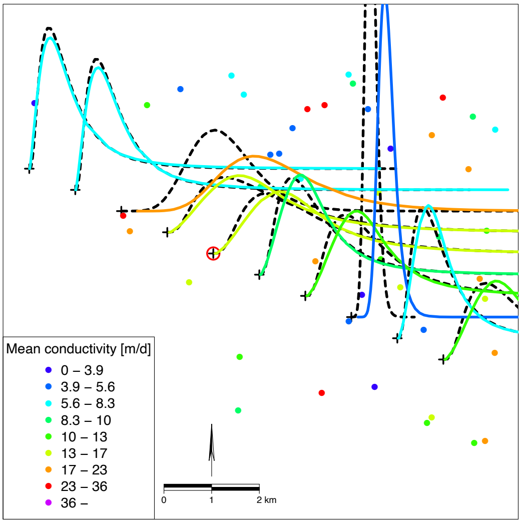

Figure 2.8: *Map with result of the kriging interpolation of conductivities. The black dashed lines show the result of a standard ordinary kriging, the colored solid lines are the results of the new proposed method. The dots are the observation locations where the color indicates the mean value. The plus signs are the kriging locations.*

the linear functions. Another difference between the two methods is the possibility to describe functions with an infinite domain. The piecewise linear method has to truncate the infinite tails at some finite value, the method of Jaroszewicz and Korzeń is able to support infinite domains by use of exponential tails.

As an example, the summation of ten standard Gaussian distributed RVs is performed. The analytical mean and variance are 0 and 10, respectively. The result of the method of Jaroszewicz and Korzeń is about 1.2178e-15 and 10 (with 14 trailing zeros), and the result of the piecewise linear method is 5.879e-5 and 10.1049. The piecewise linear PDFs are discretized with 50 bins and truncated at five times the standard deviation.

The higher accuracy is acquired at the cost of calculation time. The calculation of the transmissivity, as described by Equations (2.11) and (2.12), is used to compare the performance of both methods. In Table 2.1 the computation time is shown for the addition of one, two and three layers The calculation time of the method of

Table 2.1: *Comparison of the performance of the method of Jaroszewicz and Korzeń to the piecewise linear method.*

problem	Jaroszewicz and Korzeń	piecewise linear
	[s]	[s]
$D_1 * K_1$	1.35	0.00077
$D_1 * K_1 + D_2 * K_2$	24.1	0.0021
$D_1 * K_1 + D_2 * K_2 + D_3 * K_3$	834.	0.0033

Jaroszewicz and Korzeń is much higher than the calculation time of the piecewise linear method. Furthermore, the calculation time of the method of Jaroszewicz and Korzeń is not proportional to the number of operations but increases much more. Compared to the vertical upscaling at point scale and subsequently the horizontal interpolation in the real world example in this article, this is a very small example. In addition, in Appendix A.2 two examples from the literature containing calculations with RVs are compared to calculations using piecewise linear PDFs.

2.3 Discussion and conclusions

We developed a generic method to propagate the uncertainty of data through calculations and applied it to the upscaling of hydraulic conductivity data. The uncertain data used are represented by piecewise linear probability density functions (PDFs), which can be of any form. A similar calculation method, with a different implementation, has been described before by *Jaroszewicz and Korzeń* [2012]. However, the computation time of their method is so high that it is not easily applicable to the calculations described in this article.

Figure 2.8 shows that the magnitude of the effect of the proposed method differs between kriging locations. As may be expected, kriging locations close to observations show the largest effects on the interpolated PDFs. The results presented show a good performance of the developed PDF calculations. The implementation in upscaling of borehole data, using kriging interpolation, yields interpolated subsurface parameter data with complete PDFs instead of only the uncertainty of the mean values. Although these PDFs are a common feature of kriging, the propagation of the uncertainty of the basic data in this way throughout the calculations is new. Herewith, any distribution which can be approximated by a piecewise linear PDF can be dealt with. Compared to Monte Carlo simulation (MC), the PDF calculations yield a smoother PDF of the result. The smoothness of the result does not rely on a random number generator or the number of simulations performed.

We performed kriging on the log-values of the PDFs of the observations. This transformation relies on true log-normal distributed values when the RVs are parameterized. When the data is not exactly log-normal distributed, the back transformation of the parameters may cause a bias in the mean values. Back transformation of the PDFs does not yield a bias in mean value or variance.

Compared to calculations using parameterized PDFs or other analytical solutions, our method takes more computation time. However, we did not perform a benchmark because of the research state of the software. Nevertheless, PDF calculations can be of great value in uncertainty propagation problems where no analytical solutions are applicable. Availability of this method reduces the need for MC solutions.

Compared to the analytical PDFs, the usage of piecewise linear PDFs implies loss of accuracy in the calculated results. So care must be taken when choosing the discretization of a PDF.

3

Obtaining the most likely hydrogeological model parameter values

Abstract. Usually, subsurface data for groundwater flow models are obtained from hydrogeological models, which in turn are generated from borehole data, using upscaling techniques. Since the assumed hydraulic properties for litho-classes in boreholes are uncertain, and upscaling may add inaccuracies, the groundwater flow model has to be calibrated, and therewith the hydrogeological model. In the Netherlands, a general purpose hydrogeological model (REGIS) is developed to serve as input for multiple groundwater flow models, among other applications. For each groundwater flow model a separate hydrogeological model is derived from the REGIS model. These derived hydrogeological models are calibrated, without changing the REGIS model itself. Therewith, no direct feedback from the calibration results to the REGIS hydrogeological model is available. In this paper, a method is presented that uses a calibrated groundwater flow model to improve the quality of the general purpose hydrogeological model (layer thickness and hydraulic properties). Thereto, the uncertain layer thicknesses and conductivities are described by a joint probability density function. Subsequently, the calibrated data is used to find the most likely combination of parameters within this joint distribution. We illustrate the proposed method to a case where aquitard thickness and vertical hydraulic conductivity are estimated. In order to make the problem tractable, computationally feasible, and avoid assumptions about the distribution form, piecewise linear probability density functions are used, instead of parameterized functions.

KNOWLEDGE OF THE SUBSURFACE is of great importance to various areas of interest, like drinking water supply and heat and cold storage, among others. Characterization and modeling of the subsurface is inevitable, and geological and hydrogeological models are widely developed. The quality or uncertainty of these models depends highly on the available data and is worldwide a matter of continuous concern.

In the Netherlands, a nation-wide digital geological model (DGM) [*Gunnink et al.*, 2013] of the subsurface is constructed and maintained by the Geological Survey of the Netherlands (TNO-GSN). This three-dimensional model displays the geological units, based on the lithostratigraphical classification. The definition of the units is based on the depositional environment (e.g. marine or fluvial) and the age of the depositions. The geological units mainly coincide with formations.

The DGM is subsequently used as a framework to define the nation-wide general purpose hydrogeological model REGIS [*Vernes et al.*, 2005; *Vernes and van Doorn*, 2006]. REGIS is also developed and maintained by TNO-GSN. Within the geological units of the DGM, multiple litho-classes are recognized. Such a litho-class is a combination of the geological unit (Formation) and the lithology of the sediments (clay, sandy clay, fine sand, coarse sand, peat, etc.). Based on the hydraulic properties of the litho-classes, layers of several litho-classes are aggregated to hydrogeological units. The hydrogeological units are divided into layers with high conductivity (aquifers) and low conductivity (aquitards). So, within one geological unit, several hydrogeological units may be recognized. In REGIS, over one-hundred hydrogeological units are defined. Due to lateral differences in the geological processes, not every geological and hydrogeological unit is present everywhere in the subsurface of the Netherlands. Nevertheless, to make of REGIS a consistent hydrogeological model, all hydrogeological units are defined everywhere in the model but do have a zero thickness where absent.

All hydrogeological units are recognized and defined at the available boreholes. At the borehole locations, for each hydrogeological unit the layer thickness and the average horizontal and vertical conductivity are defined. Subsequently, these properties of all hydrogeological units are interpolated and upscaled to a grid with a resolution of $100\,\mathrm{m} \times 100\,\mathrm{m}$. The interpolation and upscaling is guided using geological knowledge about, for instance, geological processes and presence of faults. The upscaling of hydrogeological data is a process of major importance and has to be applied carefully. A vast amount of literature is available on this topic [e.g., *Dagan*, 1986; *Nœtinger et al.*, 2005; *Sanchez-Vila et al.*, 2006; *Fiori et al.*, 2011]. In the proposed method as described in this paper, the REGIS hydrogeological model is regarded as en existing model and used at the scale it is designed for. Therefore, the upscaling as implemented in the development of REGIS will not be discussed in this paper.

The grid data of REGIS define a general hydrogeological model which serves as input for multiple groundwater flow models [e.g., *Snepvangers et al.*, 2008; *Lange et al.*, 2014; *de Lange and Borren*, 2014]. Depending on the location and the extent of a groundwater flow model, only a certain subset of all hydrogeological units is present in the subsurface. Therefore, a tailor-made hydrogeological model is derived from the REGIS hydrogeological model which suits the needs of the specific groundwater flow model, hereafter denoted as the derived hydrogeological model. So multiple adjacent hydrogeological units of the REGIS model with similar hydraulic properties are aggregated to one aquifer or aquitard in the groundwater flow model. Such a derived hydrogeological model, i.e. the aquifers and aquitards of the groundwater flow model, is different from the REGIS hydrogeological model, although the total transmissivity (horizontal) and vertical resistance are equal in both models. The groundwater flow models are built and calibrated by parties as engineering consultancies or research institutes, but usually not by TNO-GSN.

It is common practice in the Netherlands to calibrate the parameters of the derived hydrogeological model without changing the hydrogeological model REGIS. Calibration is generally based on the comparison of hydraulic heads simulated by the groundwater model with observed heads in observation wells. Usually, a program, like MODFLOW [*McDonald and Harbaugh*, 1988], is used with only the transmissivity and the vertical resistance as hydraulic subsurface parameters. Therefore, calibration takes only place on these parameters of the derived hydrogeological model and not on the layer thickness and the conductivity of the hydrogeological units as stored in REGIS. Thus, no formal feedback exists between the groundwater flow model calibration results and the a priori hydrogeological parameters present in REGIS.

Any inconsistencies between the calibrated groundwater flow parameters and REGIS parameterization that occur during the calibration process are only communicated from the groundwater modelers to the (hydro)geologists on an ad hoc basis. In this paper, we introduce a more objective method to perform this communication. The proposed method makes use of the calibrated parameters of the groundwater flow model and translates these improved data back to the hydrogeological model REGIS. This yields the most likely layer thickness and conductivity of each litho-class at each grid cell, given the geological borehole descriptions and the calibration results (i.e. implicitly the head observations).

All models are to some extent uncertain or erroneous. Important sources of uncertainty are errors in the schematization, identification and the parameter values and observations. This uncertainty, and the lack of an exhaustive number of observations, yields the possibility of different models with different parameterization and schematization but with comparable performance (equifinality) with respect to the observations. In the literature, the concepts of equifinality [*Beven and Binley*, 1992; *Beven*, 2006; *Efstratiadis and Koutsoyiannis*, 2010], and multi-objective calibration [e.g., *Gupta et al.*, 1998; *Singh et al.*, 2008; *Efstratiadis and Koutsoyiannis*, 2010] are used to reflect the model uncertainty with multiple instances of the same model. Currently, only one parameterization of the most recent version of the REGIS hydrogeological model is made available, together with some uncertainty information. It is up to the groundwater flow modeler to decide the necessity of multiple instances of the derived hydrogeological model in the calibration process. The method proposed in this paper is described for one instance of a calibrated groundwater flow model, but may be used for multiple instances as well to find a distribution of most likely parameter values.

This paper is organized as follows. In Section 3.1 the methodology is described, which, in this paper, focuses on the hydraulic resistance of aquitards. Here, Section 3.1.1 describes the core part of the proposed method. In Section 3.2 the study area and the data used are presented. The results are presented in Section 3.3. The applicability of the method, and the interaction with the calibration of a groundwater flow model are discussed in Section 3.4. In Section 3.5 conclusions are drawn and an outlook for further research is given.

3.1 Methodology

The proposed update method is an addition to the prevailing modeling practice with groundwater flow models derived from the REGIS hydrogeological model. In Figure 3.1 the implementation of the update method in the current modeling process is depicted. In the top row of this figure, the REGIS hydrogeological model serves as input to multiple groundwater flow models, but without a formal feedback of the calibrated results to the REGIS model. This gap in the modeling process can be filled in by the update algorithm (bottom row in Figure 3.1). This algorithm yields an updated version of REGIS, which is applicable in new studies. The methodology described in this section consists of several steps. The flowchart in Figure 3.2 shows the four processing steps and the three sources of input data. The core part of the method is step 4 (Section 3.1.1), which is described first. Herein, for each grid cell for each litho-class l the joint distribution with the marginal distributions of the layer thickness (D_l) and the vertical hydraulic conductivity (K_l) is defined. From this joint distribution, the most likely parameter values (layer thickness and conductivity) of all litho-classes are found conditional on the calibrated vertical resistance (c_m) of a groundwater flow model. Here, the most likely parameter combination is the combination with the highest joint probability density.

To be able to build the joint distribution by using the distributions of D_l and K_l as the marginal distributions, these marginal distributions need to be known. In the

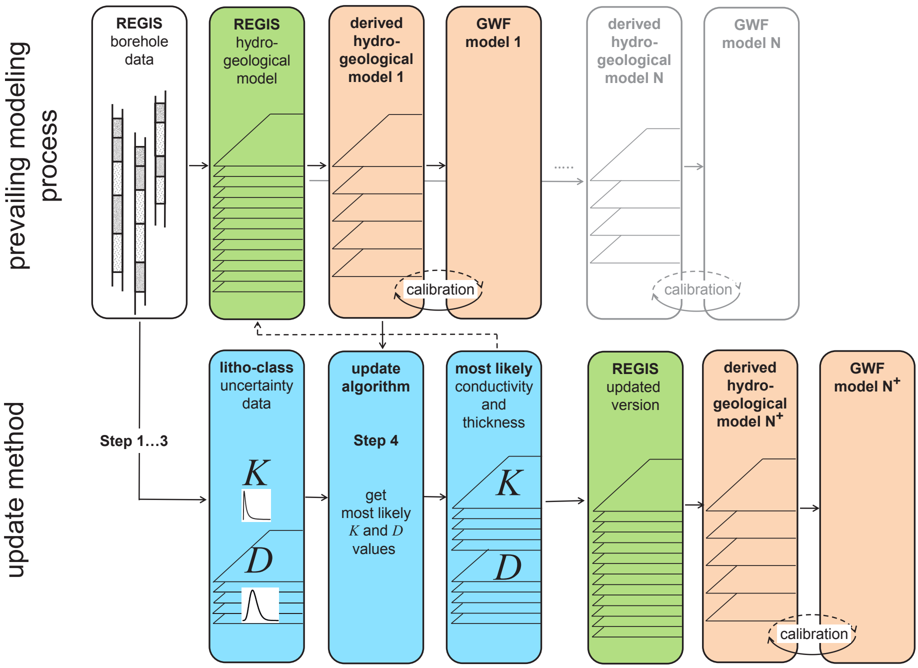
prevailing modeling process
update method
REGIS borehole data
REGIS hydro-geological model
derived hydro-geological model 1
GWF model 1
derived hydro-geological model N
GWF model N
calibration
calibration
Step 1...3
litho-class uncertainty data
K
D
update algorithm
Step 4
get most likely K and D values
most likely conductivity and thickness
K
D
REGIS updated version
derived hydro-geological model N+
GWF model N+
calibration

Figure 3.1: *Relation between the prevailing modeling process with* REGIS *(top row) and the update method of this paper (bottom row). The step numbers refer to the steps in Figure 3.2 and the descriptions in Section 3.1. The dashed arrow shows a feedback in case of detected model identification errors (see discussion in Section 3.4).*

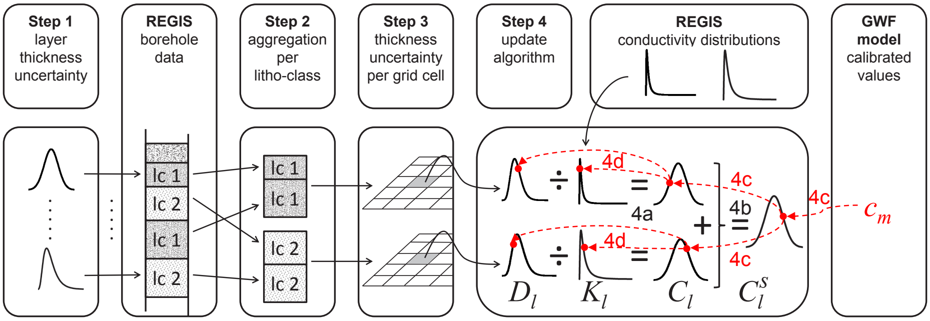
Step 1 layer thickness uncertainty
REGIS borehole data
Step 2 aggregation per litho-class
Step 3 thickness uncertainty per grid cell
Step 4 update algorithm
REGIS conductivity distributions
GWF model calibrated values
lc 1
lc 2
lc 1
lc 2
lc 1
lc 1
lc 2
lc 2
4d
4a
4d
4c
4b
4c
4c
c_m
D_l
K_l
C_l
C_l^s

Figure 3.2: *Flowchart of the method as described in Section 3.1. The descriptions of step 1, 2, 3 and 4 are found in Sections 3.1.2, 3.1.3, 3.1.4, and 3.1.1, respectively. This chart shows an example with two litho-classes (lc 1 and lc 2) in one aquitard.*

REGIS system, the distributions of the conductivities are defined for each litho-class. However, no information about the uncertainty of the layer thicknesses is readily available. Therefore, in step 1 (Section 3.1.2), a probability density function (PDF) of the thickness is assigned to each defined litho-layer from the interpreted borehole data from REGIS. In step 2 (Section 3.1.3), these data are aggregated to one PDF of the total thickness of each litho-class at each borehole. In step 3 (Section 3.1.4), a kriging interpolation of the litho-layer thickness is performed, using the result of step 2 as observations. This yields a PDF for the thickness of each litho-class at each grid cell, necessary as input for the algorithm in step 4 (Section 3.1.1). The resolution of this grid is comparable to the resolution of the calibrated groundwater flow model.

3.1.1 Update algorithm

The vertical hydraulic resistance of a litho-layer can be derived from observations of the layer thickness and the vertical conductivity of the deposits. These observations always yield uncertain parameter values and they might not be representative for the required model scale. Instead, the calibrated parameters of the derived hydrogeological model of the groundwater flow model are used as input for the update algorithm (Figure 3.1). The uncertain parameters (litho-layer thickness and conductivity) are treated as random variables described by their probability density functions (PDFs). In this paper, all random variables (RVs) are described by piecewise linear PDFs [*Kaczynski et al.*, 2012; *Vander Wielen and Vander Wielen*, 2015] from which all calculations can be performed independent of the type of distribution assumed. Performing elementary operations and kriging interpolation with piecewise linear PDFs is described in *Lourens and van Geer* [2016].

Let the value of the vertical resistance of an aquitard at grid cell u, denoted by $c_{\mathrm{m}}(u)$, be the result of the calibration of a groundwater flow model. This calibrated resistance is assumed to be the true value. In the proposed method, no uncertainty of the calibrated groundwater flow model is included, so $c_{\mathrm{m}}(u)$ is treated as a deterministic parameter. Furthermore, in accordance with the REGIS assumptions, we assume that the PDF of the hydraulic conductivity for a given litho-layer does not change in space.

The vertical resistance of a litho-layer is calculated as

$$c_l(u) = d_l(u)/k_l(u), \tag{3.1}$$

where $c_l(u) \in C_l(u)$ is the vertical hydraulic resistance, $d_l(u) \in D_l(u)$ is the layer thickness, $k_l(u) \in K_l$ the vertical hydraulic conductivity, l the litho-class, and u denotes the location. Although K_l is assumed to be location independent, k_l certainly not. The variables C_l, D_l, and K_l are RVs. The PDFs of D_l and K_l are $f_{D_l}(d_l)$ and $f_{K_l}(k_l)$, respectively. The variables D and K are assumed to be statistically independent. Hereafter, for readability the location indicator (u) is dropped from the equations.

The total resistance of the litho-layers 1 until l is calculated as the summation of the individual resistances, as

$$c_l^s = \sum_{i=1}^{l} c_i = c_{l-1}^s + c_l \quad \text{for} \quad 2 \leq l \leq n, \tag{3.2}$$

where n is the number of litho-classes in the aquitard, $c_l^s \in C_l^s$, and $C_1^s = C_1$. The superscript s denotes the summation of the vertical resistances from litho-layer 1 up to layer l. Hence, $c_n^s \in C_n^s$ is the total aquitard resistance.

With the expression of Equation (3.1), the joint PDF $f_{C_l}(\theta_l)$ of D_l and K_l can be written as

$$f_{C_l}(\theta_l) = f_{D_l}(d_l) f_{K_l}(k_l), \tag{3.3}$$

with $\theta_l = (d_l, k_l)$. Note that $f_{C_l}(\cdot)$ represents the joint PDF of D_l and K_l, and does not represent the PDF of C_l. The $2n$-dimensional joint distribution of C_n^s is defined by

$$f_C(\theta) = \prod_{l=1}^{n} f_{C_l}(\theta_l), \tag{3.4}$$

where $\theta = (\theta_1, \ldots, \theta_n)$, and the index C denotes the total aquitard resistance.

With the proposed method, we search for the most likely values of θ given the observation c_{m}. The most likely value is defined as the value with the highest probability density, or mode, of $f_C(\theta)$ conditional on $\sum_{l=1}^{n} c_l = c_{\mathrm{m}}$. This conditional density function is written as

$$f_C(\theta | c_{\mathrm{m}}) \propto \prod_{l=1}^{n} f_{C_l}(\theta_l) \quad \text{for} \quad \sum_{l=1}^{n} c_l = c_{\mathrm{m}}. \tag{3.5}$$

In this equation, the normalizing constant is left out. Since $f_C(\theta|c_{\mathrm{m}})$ is only used to find a conditional mode, this expression suffices. So the $\propto$ sign is interpreted as an equality sign. Herewith, the maximization function of θ is defined as

$$\vartheta_C(c_{\mathrm{m}}) = \underset{\theta}{\operatorname{argmax}} \, f_C(\theta | c_{\mathrm{m}}). \tag{3.6}$$

This function returns for observation c_{m} the most likely parameter values of θ, denoted by $\hat{\theta}$. Since θ is of size $2n$, and all marginal PDFs are described as piecewise linear functions, it is infeasible to write Equation (3.6) as a workable analytical expression. Therefore, a stepwise method is derived to find the mode of $f_C(\theta|c_{\mathrm{m}})$. Equivalently to Equations (3.5) and (3.6), for each litho-layer the conditional density function

$$f_{C_l}(\theta_l | c_l) \propto f_{C_l}(\theta_l) \quad \text{for} \quad d_l / k_l = c_l, \tag{3.7}$$

and the maximization function

$$\vartheta_{C_l}(c_l) = \underset{\theta_l}{\operatorname{argmax}} \, f_{C_l}(\theta_l | c_l) \tag{3.8}$$

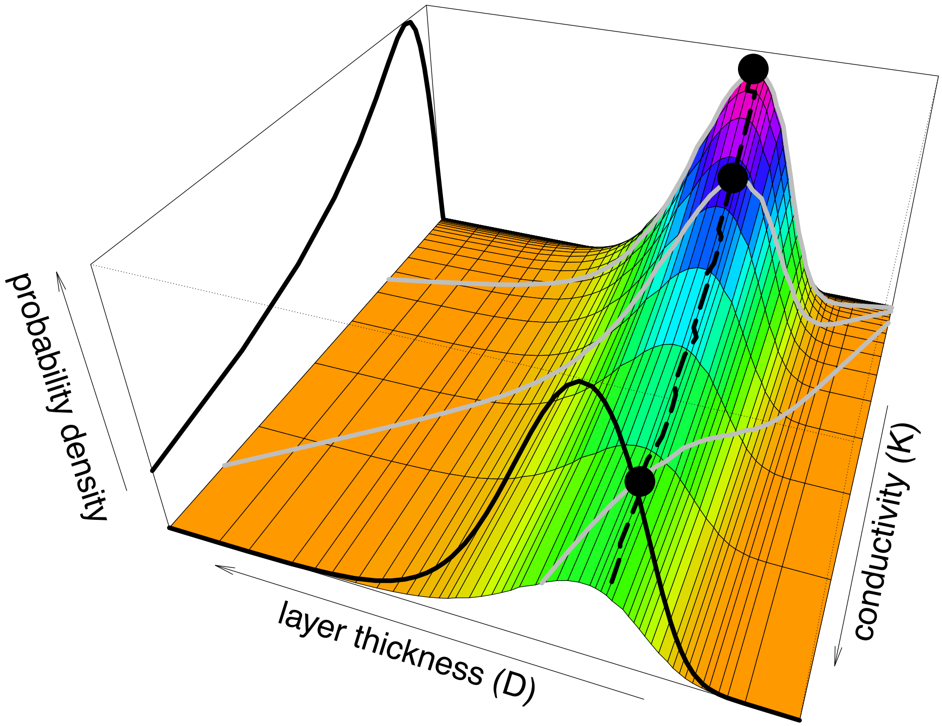

Figure 3.3: Joint PDF of the litho-layer thickness and the conductivity. The gray lines denote equi C-lines. The black dashed line connects the maximum density points of each value of C. The two marginal distributions are shown at the side-panes.

are defined. Hereafter, the steps 4a until 4d refer to the steps in the flowchart in Figure 3.2.

Step 4a is the derivation of the maximization function $\vartheta_{C_l}(c_l)$ for a joint PDF with the two marginal distributions $f_{D_l}(\cdot)$ and $f_{K_l}(\cdot)$. To illustrate this, Figure 3.3 shows the joint distribution $f_{C_l}(\theta_l)$. At the horizontal axes, the layer thickness (D_l) and the conductivity (K_l) are shown with their respective marginal PDFs. The vertical axis shows the joint probability density. The three solid gray lines are, as an example, drawn for three different values of c_l. Each line shows the conditional density function for a given value of c_l, i.e. $f_{C_l}(\theta_l|c_l)$. The mode of each conditional density function in Figure 3.3, is denoted by a black dot. The maximization function $\vartheta_{C_l}(c_l)$ returns the marginal values of this point (θ_l). The derivation of the maximization functions, using piecewise linear marginal PDFs, is given in Appendix B.1. We call the function connecting the modes for all possible values of c_l the mode density function (MDF), which is defined as

$$h_{C_l}(c_l) = f_{C_l}(\theta_l = \vartheta_{C_l}(c_l)). \tag{3.9}$$

The mode values are calculated for a number of values of c_l, which yields a piecewise linear MDF. In Figures 3.3 and 3.4, this MDF is denoted by a dashed black line. In Figure 3.4, the update algorithm is depicted for n litho-layers. The MDF returns the probability density for a reduced universe [*Bolstad*, 2007, p. 62] and is therewith not a PDF since it not necessarily integrates to 1. Nevertheless, it can be used as a marginal distribution in the next step. This is depicted with the black arrows in Figure 3.4.

The former step yields an n-dimensional joint PDF, with as marginal distributions the MDFs $h_{C_l}(c_l)$ for $l = 1, \dots, n$, which still may be too large to be evaluated at once. Step 4b is similar to the first step, but instead of a division (d_l/k_l), a summation as in Equation (3.2) ($c^s_{l-1} + c_l$) is performed. This is depicted in the right column in Figure 3.4. Equivalently to step 4a, the next sets of functions are defined and evaluated

$$\left.\begin{aligned} f_{C^s_l}(\theta^s_l) &= h_{C^s_{l-1}}(c^s_{l-1}) h_{C_l}(c_l) \\ \vartheta_{C^s_l}(c^s_l) &= \underset{\theta^s_l}{\operatorname{argmax}}\, f_{C^s_l}(\theta^s_l | c^s_l) \\ h_{C^s_l}(c^s_l) &= f_{C^s_l}(\theta^s_l = \vartheta_{C^s_l}(c^s_l)), \end{aligned}\right\} \quad \text{for} \quad l = 2, \dots, n \tag{3.10}$$

where $C^s_1 = C_1$, $h_{C^s_1}(c^s_1) = h_{C_1}(c_1)$, and $\theta^s_l = (c^s_{l-1}, c_l)$. For all maximization functions $\vartheta_{C^s_l}(c^s_l)$ the argument value of c^s_l is unknown. However, for the last maximization function $\vartheta_{C^s_n}(c^s_n)$ the argument value is available with $c^s_n = c_\mathrm{m}$ being the total calibrated vertical resistance of the aquitard.

In the next step, 4c, all maximization functions $\vartheta_\cdot(\cdot)$ are used to find the mode estimates $\hat{\theta}$ for all marginal distributions, given c_m, starting with $\hat{\theta}^s_n = \vartheta_{C^s_n}(c_\mathrm{m})$, which yields $\hat{\theta}^s_n = (\hat{c}^s_{n-1}, \hat{c}_n)$. This is shown in the top-right pane of Figure 3.4. Herewith, $\hat{\theta}^s_n$ contains the mode estimates of the total vertical resistance of litho-layers $1 \dots n-1$, i.e. $\hat{c}^s_{n-1}$, and the vertical resistance of litho-layer n, i.e. $\hat{c}_n$. In general, with $\hat{c}^s_n$ being known, subsequent evaluation of

$$\hat{\theta}^s_l = \vartheta_{C^s_l}(\hat{c}^s_l) \quad \text{for} \quad l = n, \dots, 2, \tag{3.11}$$

yields $\hat{\theta}^s_l = (\hat{c}^s_{l-1}, \hat{c}_l)$. This is in Figure 3.4 depicted with the red arrows. Herewith, all values $\hat{c}_l$ are available.

In the last step, 4d, the values $\hat{c}_l$ are used to find the mode values of d_l and k_l. For $l = 1, \dots, n$, the expression $\hat{\theta}_l = \vartheta_{C_l}(\hat{c}_l)$ yields $\hat{\theta}_l = (\hat{d}_l, \hat{k}_l)$. Herewith, all conditional mode estimates of the marginal distributions $f_{D_l}(\cdot)$ and $f_{K_l}(\cdot)$ are known.

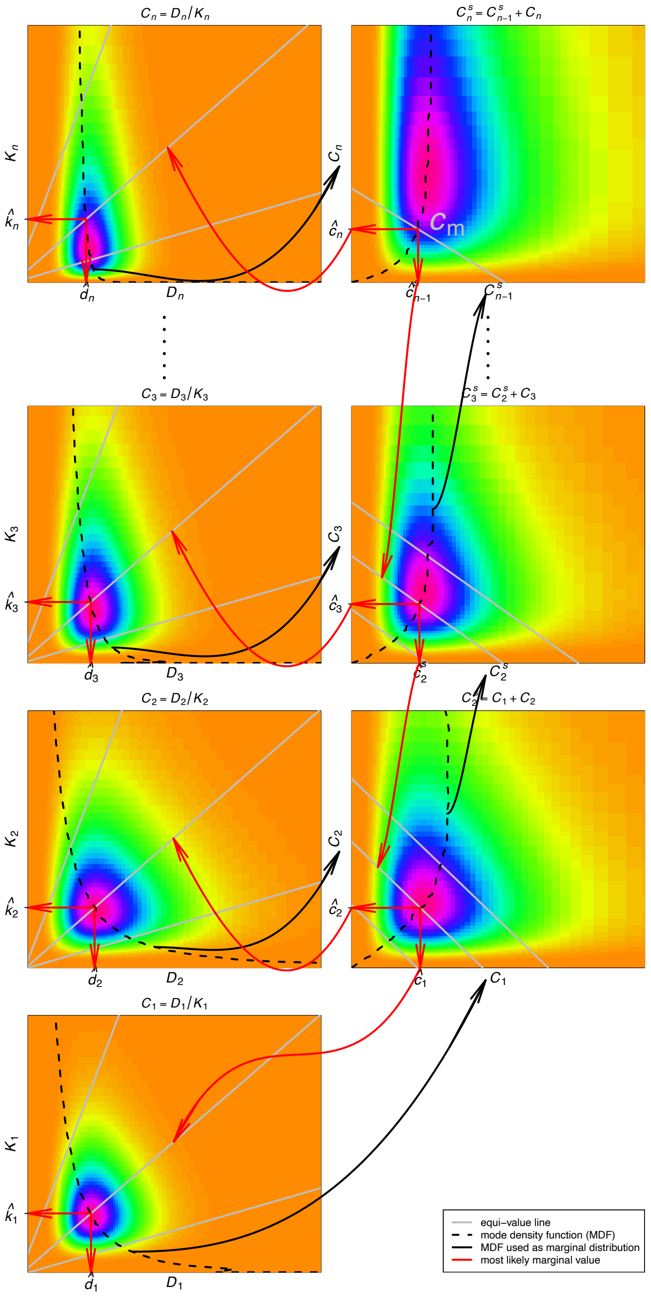

Figure 3.4: *Scheme of methodology step 4. Every graph shows a joint distribution of two marginal distributions as in Figure 3.3. Firstly, MDFs are generated from D and K (left column). Secondly, the MDFs are used as marginal distributions of C (right, black arrows), and new MDFs are generated. Thirdly, in the top right graph, the observation C_m is used to find the most likely marginal values (intersection of equi-value line with MDF). Finally, all most likely marginal values of C, D and K are found (red arrows).*

With this method, a multidimensional joint PDF can successively be evaluated to find the most likely conductivity and layer thickness values of all marginal distributions, conditional on an observation of the total aquitard resistance.

3.1.2 Layer thickness uncertainty

This section describes step 1 of the flowchart in Figure 3.2. The method presented in the former section needs a quantification of the uncertainty of litho-layer thicknesses. However, quantitative data about this uncertainty are usually not available. In this section, a method is described to provide all litho-layers of the borehole descriptions with an appropriate uncertainty of the layer thickness. There are certainly more sources of error in borehole descriptions, like misclassification of sand and clay layers during drilling, or the interpretations of the litho-class by the hydrogeologist, but in this study we only evaluate the possible error in the layer thickness.

During drilling of a borehole, the measured layer thicknesses are always rounded off. This causes uncertainty in the layer thickness observations. The magnitude of the round-off error depends, among others, on the drilling method and the way the borehole descriptions are made. Therefore, it is likely that drilling methods which can distinguish the layers more accurately have a smaller round-off error than drilling methods with a lower accuracy. Reversing this reasoning it may be concluded that small round-off values give a more accurate layer thickness than large round-off values. The question is how to recognize the order of magnitude of a round-off error in the borehole description data, and what may be an appropriate uncertainty to ascribe to a specific round-off error.

From the REGIS data base, about 475 000 litho-layer thicknesses of about 16 000 borehole descriptions are available. The remainder of all these thicknesses, when dividing by one meter, is calculated and shown as a cumulative distribution in Figure 3.5. From this figure, it can be seen that round off to one meter (remainder is 0) is done very often (44 %). Also round off to fifty (12 %), ten (30 %) and five (8 %) centimeters is done more often than to one (6 %) centimeter. Truncation to a smaller value than one centimeter is not stored in the data base. The number of layers in each truncation class is counted after removing the layers counted in a higher truncation class (a class with a higher round-off value). Obviously, values of layers of a lower class often coincide with values of a higher class. This makes the above counting biased. Although this can be statistically corrected for the distribution as a whole, it can not easily be corrected for the individual layers. Therefore, no correction is applied and this error is accepted in the described method.

Since no quantitative information is available about the uncertainty of the litho-layer thicknesses, an arbitrary choice has to be made. This choice implies the type of PDF and the magnitude of variance. To justify the choice, a sensitivity analysis has been carried out to test the performance of different options. These options include two types of distributions, and several magnitudes of the variances. When the type of distribution is unknown, the Gaussian distribution is usually a safe choice,

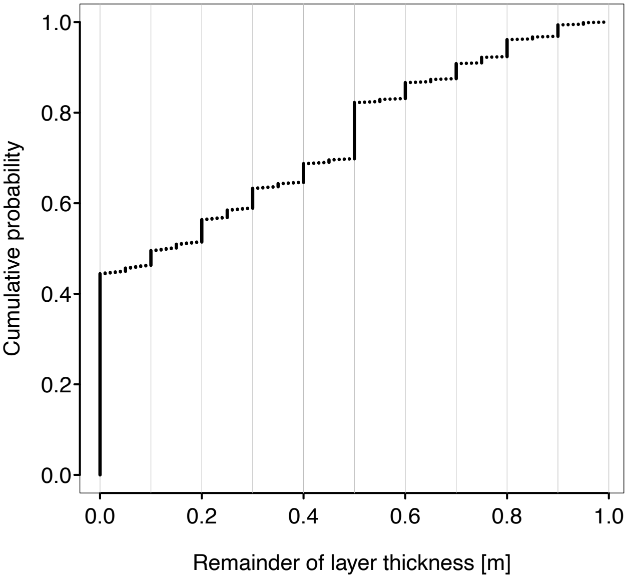

Figure 3.5: *Cumulative distribution of the remainder of about 475 000 litho-layer thicknesses after division by one meter. The vertical lines show the position of the round-off values at every ten centimeters.*

Table 3.1: *Three classes of standard deviation related to the truncation classes.*

truncation class [m]	low [m]	medium [m]	high [m]
0.01	0.002	0.01	0.05
0.05	0.010	0.05	0.25
0.10	0.020	0.10	0.50
0.50	0.100	0.50	2.50
1.00	0.200	1.00	5.00

because the round-off errors may be assumed symmetric around zero. Since the Gaussian distribution may yield negative layer thicknesses, the log-normal distribution has been tested as well. We have chosen to relate the standard deviation for each litho-layer thickness linearly to its truncation class value (Table 3.1). For the low standard deviation class this factor is $1/5$, for the medium class 1, and for the high class 5.

When a litho-class is observed absent in a borehole, it should get a thickness of 0 m. However, an expected thickness value of zero yields problems with the assignment of a PDF to this observation. When a Gaussian distribution is chosen, the layer thickness will be less then zero with a probability of 0.5. When choosing a log-normal distribution it is impossible to assign a variance greater than zero to the observation. Therefore, a small positive value has to be chosen for the 0-thickness observations. The choice of an appropriate value is described in Section 3.3.3.

3.1.3 Data preparation

The described method needs complete borehole descriptions for a model layer (aquifer or aquitard) at all borehole locations. Therefore, when a description in a borehole is incomplete for a model layer, this borehole is neglected for that layer.

Within the extent of a study area and within the considered model layer, a limited set of litho-classes is found. Because of heterogeneity of the subsurface, not every litho-class is present at every borehole location. However, the absence of a litho-class in a certain borehole is an observation as well. Therefore, when a litho-class is absent in a borehole, it is added with a zero layer thickness. The assignment of the variance to the layer thickness has been described in Section 3.1.2.

A litho-class may appear multiple times within one model layer in one borehole. The thicknesses and variances of all these occurrences are added to one thickness and variance before further processing (step 2 of the flowchart in Figure 3.2). Consequently, the horizontal connectivity of individual litho-layers of a litho-class between boreholes is neglected. The result of this processing step is a PDF of the layer thickness for each litho-class at each borehole location. The term litho-layer is used for the individual litho-layers as well as for the aggregated litho-layers.

3.1.4 Assessment of layer thickness uncertainty per grid cell

The algorithm, as described in Section 3.1.1, needs for each litho-class the PDFs of the layer thickness ($D_l(u)$) and the hydraulic conductivity (K_l) at each grid cell. This section describes the assessment of the PDF of $D_l(u)$, which is step 3 of the flowchart in Figure 3.2.

In the information system REGIS, a PDF of the hydraulic conductivity is assigned to each litho-class, independent of the spatial coordinates. So everywhere in the subsurface where a particular litho-class exists, the probability distribution of the hydraulic conductivity is assumed to be known. Therefore, only the PDFs of the layer thickness for each litho-class have to be spatially predicted. This spatial prediction is performed by using ordinary kriging (OK). For every litho-class a semi-variogram model for the litho-layer thickness is estimated. Since layer thicknesses are greater than or equal to zero, the interpolation method must account for this [*Tolosana-Delgado and Pawlowsky-Glahn*, 2007]. In Section 3.3.4, several interpolation options are evaluated to select the most appropriate ones, concerning the observed data.

Using kriging interpolation, the estimation of the interpolated thickness $\hat{D}(u_0)$ at the unobserved location u_0 writes [*Isaaks and Srivastava*, 1989, p. 282]

$$\hat{D}(u_0) = \sum_{\alpha=1}^{m} \lambda_\alpha D(u_\alpha), \tag{3.12}$$

where m is the number of observations, λ_α are the kriging weight factors, and $D(u_\alpha)$ are the observations at the locations u_α. Usually, the observations $D(u_\alpha)$ are treated as deterministic values. In this study, $D(u_\alpha)$ is the complete PDF of the layer thicknesses, described as a piecewise linear function. This method yields a PDF of the interpolated litho-layer thickness $\hat{D}(u_0)$. Subsequently, the PDF of the interpolation $(\hat{D}(u_0))$ and the PDF of the interpolation error are added to achieve a PDF containing all uncertainties. In previous work, this method is described in detail [*Lourens and van Geer*, 2016]. Generation of the PDF of the litho-layer thickness of the observations, and the choice of its attributes, like variance and shape, is described in Section 3.1.2

Ordinary kriging tends to generate negative weight factors, beside the positive ones, when the spatial distribution of the observations is somehow unbalanced around the estimation location, known as the screen effect. Apart from the physical meaning of negative weight factors, this screen effect influences the interpolated result [*Goovaerts*, 1997, p. 176]. To avoid this, the kriging algorithm is modified following the method as described by *Deutsch* [1996]. Herewith, the observation location with the most negative weight factor is removed from the subset of locations for the current kriging location. This is repeated until no negative weight factors are calculated anymore, or until less than the required minimum number of observations is reached. In the latter case, a missing value is assigned to the corresponding kriged location.

Subsequently, the PDFs of the layer thickness and the PDFs of the hydraulic conductivity serve as input for the update algorithm of Section 3.1.1.

3.2 Study area and available data

The method developed in this paper has been tested and evaluated using a case study. The location of the study area is shown in Figure 3.6. The size of this area is 20 km×25 km with a grid size of 100 m × 100 m. The used borehole data originate from borehole descriptions as administered by the Geological Survey of the Netherlands (TNO-GSN), and the hydrogeological interpretations as stored in the REGIS information system. In REGIS, litho-classes are assigned to all identified layers from every borehole. The definition of the litho-classes is based on lithological properties and lithostratigraphical units. Each litho-class is provided with two probability density functions (PDFs) of the hydraulic conductivity, one for the horizontal conductivity and one for the vertical conductivity. Subsequently, these litho-classes are aggregated to hydrogeological units. These data are thus suitable to be used in numerical groundwater flow models. The data include layer depths, litho-classes, and hydrogeological units.

The calibrated layer properties (transmissivity and hydraulic resistance values) originate from the AZURE groundwater flow model, developed by Deltares, the Netherlands [*de Lange and Borren*, 2014]. The hydrogeological model of AZURE is derived from the hydrogeological model REGIS and therefore suitable to perform

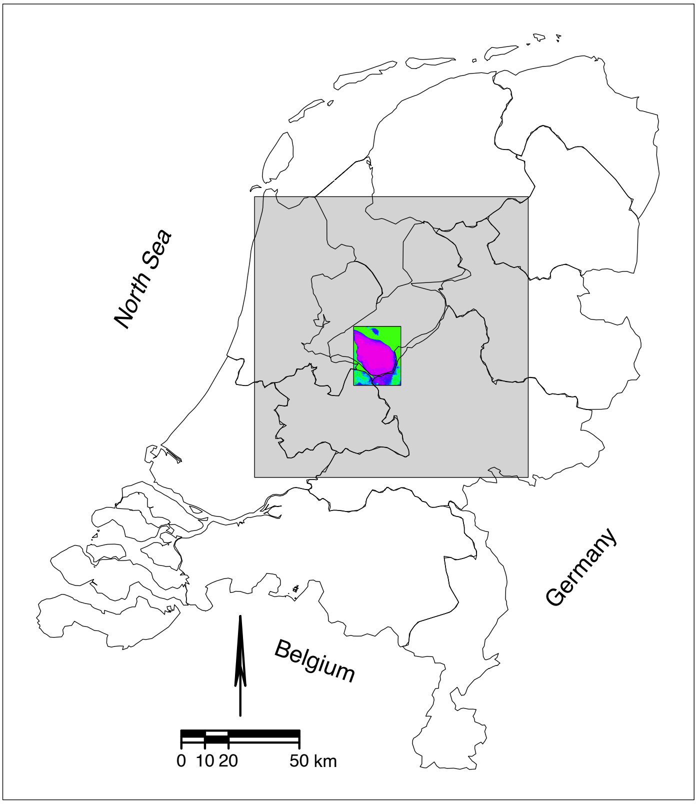

Figure 3.6: *Study area. The gray area is the extent of the AZURE groundwater flow model. The small rectangle denotes the study area with the vertical resistance, as shown in Figure 3.7, depicted.*

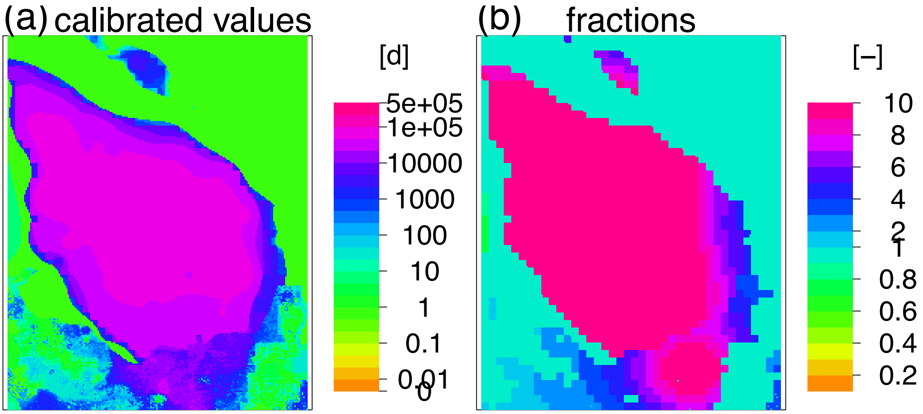

Figure 3.7: The ***(a)*** *calibrated hydraulic vertical resistance of the aquitard and* ***(b)*** *the quotient of the calibrated and the uncalibrated resistance. The majority of the calibrated resistance in the clay patch is about ten times the uncalibrated resistance.*

Table 3.2: Probability data of the vertical hydraulic conductivity values of each litho-classas provided by the REGIS information system. The distributions are defined by the 2.5 and 97.5 % percentile values and are assumed to be log-normal. The presented mean and SD are derived from the PDFs.

litho-class	2.5 %	97.5 %	mean	sd	description
	[m d^{-1}]	[m d^{-1}]	[m d^{-1}]	[m d^{-1}]	
EE-k	7.3e-5	0.0219	3.64e-3	9.74e-3	Eem Fm., clay
EE-kz	7.3e-5	0.301	4.46e-2	0.372	Eem Fm., sandy clay
EE-v	6.4e-4	0.32	5.03e-2	0.166	Eem Fm., peat
EE-zf	7.3e-5	2.88	0.548	12.4	Eem Fm., fine sand
EE-zm	1.4	29.7	8.74	7.98	Eem Fm., medium sand
UR-kz	1.7e-4	0.29	4.25e-2	0.239	Urk Fm., sandy clay
UR-zg	2.4	160.7	34.9	51.2	Urk Fm., coarse sand

this study.

In this case study, we focus on the fourth aquitard in the AZURE groundwater flow model. This aquitard is a high vertical resistance clay patch, surrounded by an area where the clay layer is thin or absent. This aquitard is found between 20 and 85 m below surface level. To meet the numerical requirements of the groundwater flow model, a minimum vertical resistance of one day is used in the area where the aquitard is absent. The calibrated vertical resistance of the aquitard and the ratio calibrated/uncalibrated resistance are depicted in Figure 3.7. This ratio shows the modification of the vertical resistance by the calibration procedure.

The aquitard consists in the study area of seven different litho-classes. The hydraulic properties of these litho-classes, as defined in REGIS, are shown in Table 3.2.

Table 3.3: Variogram model for thickness of each litho-class. The range of litho-class UR-zg could not be estimated and is set to an arbitrary value.

litho-class	type	range [m]	sill [m^2]
EE-k	exponential	1800	43
EE-kz	exponential	2000	42
EE-v	exponential	4000	0.5
EE-zf	exponential	1200	6
EE-zm	exponential	800	6
UR-kz	exponential	300	5
UR-zg	exponential	(400)	4

Not all litho-classes do have characteristic properties for aquitards. The sand classes (EE-zf, EE-zm, UR-zg) have a much higher conductivity than the clay and peat classes (EE-k, EE-kz, EE-v, UR-kz). Since the deposits of these sand classes are embedded in low conductivity layers, they are part of the aquitard and modeled as such.

Table 3.3 shows the variogram models of the thickness of the litho-layers as derived from the borehole data. The accompanying experimental variograms are given in Figure 3.8. The range of the variogram model of litho-class UR-zg could not be estimated due to lack of data, and is set to an arbitrary value of 400 m. The interpolation, as described in Section 3.1.4, is performed using block kriging at a grid with 250 m wide cells and a block discretization of sixteen points. A minimum of four and a maximum of sixteen observations is used for each interpolation.

3.3 Results

3.3.1 Improved estimates of the litho-class properties

The calibrated vertical resistance of the aquitard from the groundwater flow model has been divided over the seven litho-classes, according to the method described in Section 3.1. This is shown in Figure 3.9. The major part of the vertical resistance is assigned to the EE-k and EE-kz litho-classes. As may be expected, the contributions of the litho-classes of coarser deposits to the total vertical resistance of the aquitard is small. The resistance assigned to these classes appears to be low, compared to the resistance of the clay deposits. Beside the high conductivity of the sediments, these sandy litho-classes exist only in a minority of the observations. This leads to thin litho-layers in the majority of the study area (Figure 3.10), and thus to a negligible contribution to the vertical resistance. Therefore, we focus on the two most important litho-classes EE-k (clay) and EE-kz (sandy clay).

Figure 3.11 shows the improved layer thickness of litho-classes EE-k and EE-kz

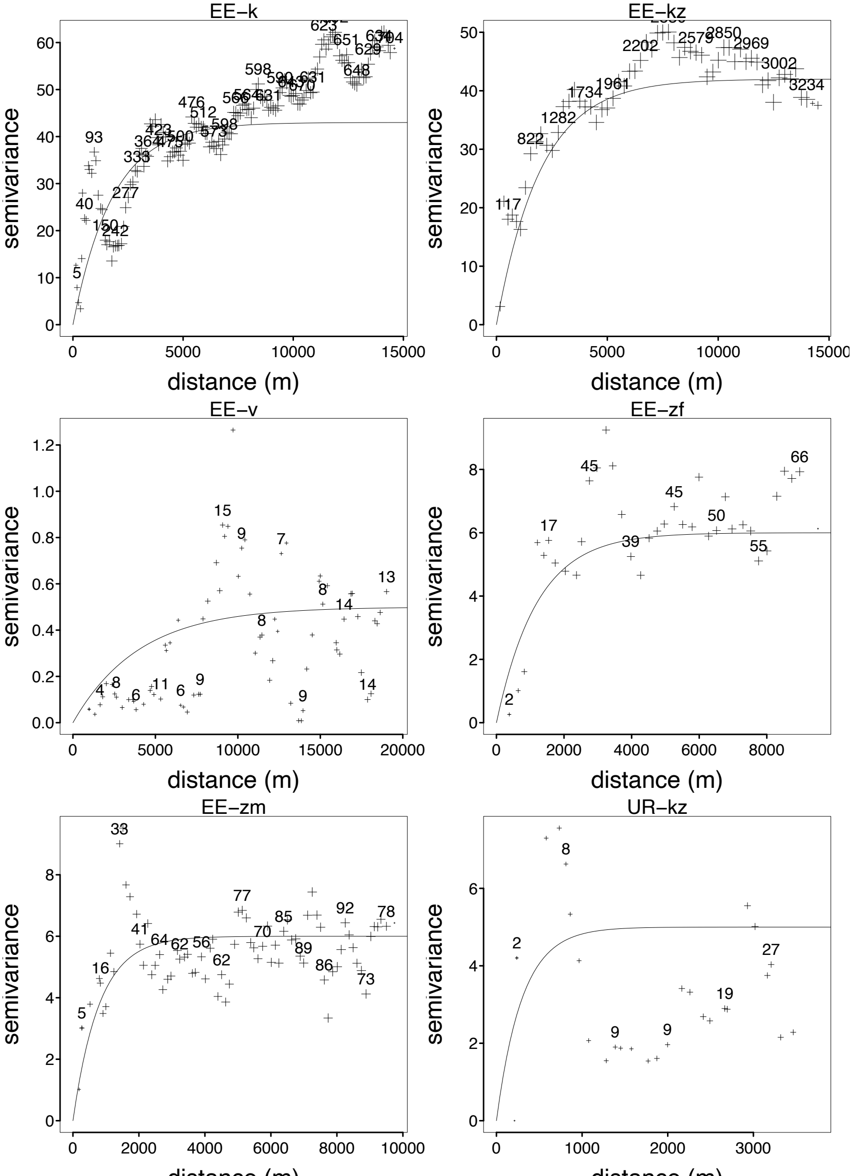

Figure 3.8: Experimental variograms and variogram models for the layer thickness of each litho-class as used in this chapter. The size of the plus signs is proportional to the number of observation pairs used to calculate the semivariance. At every fifth point the number of pairs is written. The variogram parameters are found in Table 3.3.

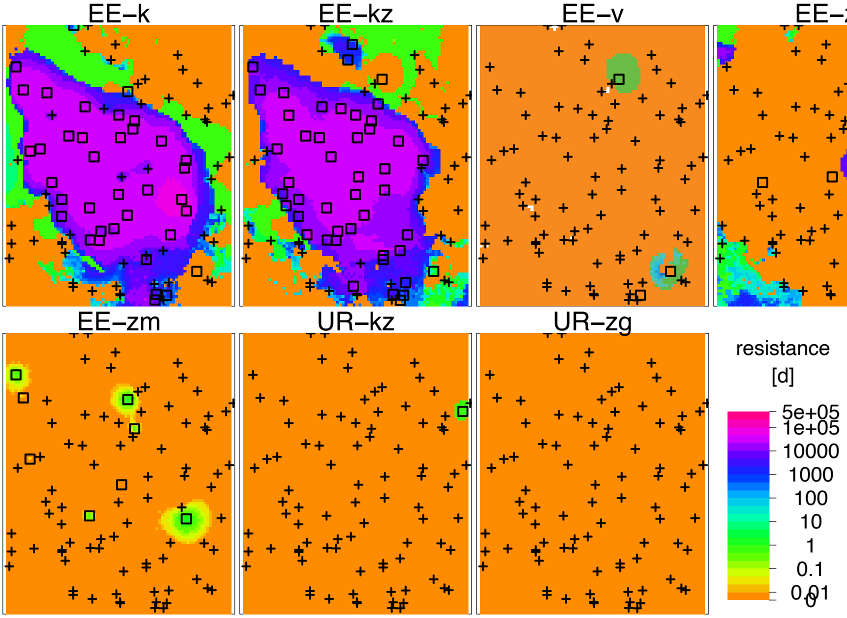

Figure 3.9: *Most likely vertical resistance. The squares denote observations where the litho-class is present, the plus signs where it is absent.*

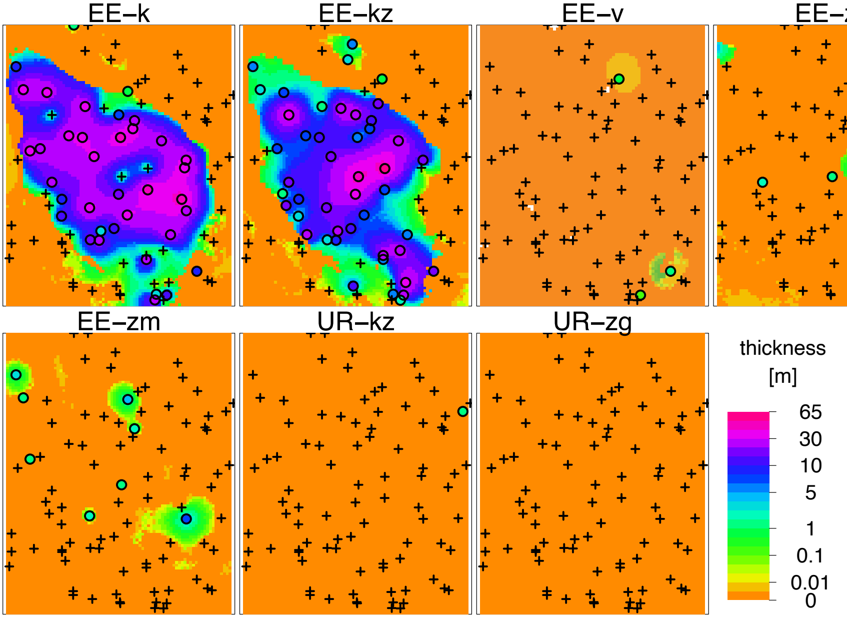

Figure 3.10: *Most likely thickness of each litho-class. The circles denote observations where the litho-class is present, the plus signs where it is absent. The circles are colored with the observed thickness.*

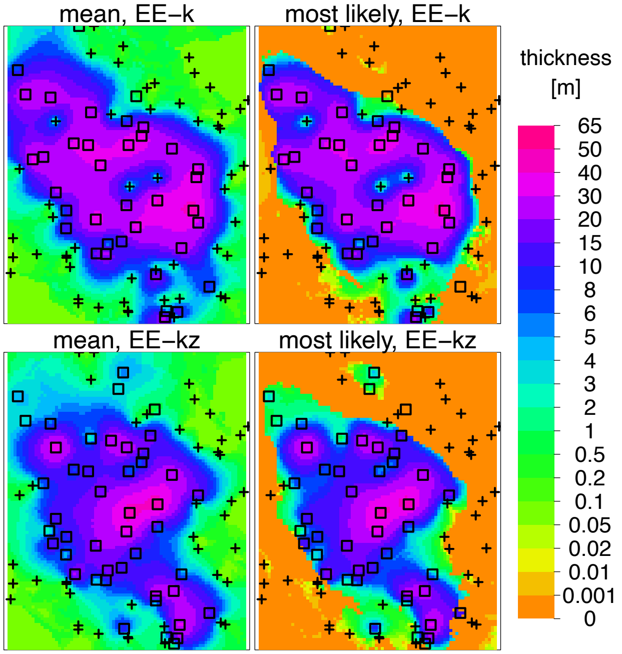

Figure 3.11: *Thickness of litho-class EE-k (top) and EE-kz (bottom). Mean kriging thickness (left) compared to most likely thickness (right).*

compared to the mean values of the PDFs of the interpolation. This mean value is the result of a kriging interpolation with deterministic valued observations, which is a common way of interpolation. As can be seen, the proposed method is able to reduce the litho-layer thickness to negligible values in the area where the aquitard is absent, whereas kriging interpolation results in more smooth patterns. The steep gradient of the layer thickness is more in agreement with the calibrated resistance in the groundwater flow model as well as the geological understanding.

The calibrated vertical resistance at a grid cell, which is used as an observation, is the total resistance of all litho-classes within the aquitard. The proposed method yields for each litho-class for each grid cell the most likely resistance, thickness and conductivity values. The position of these improved values in their a priori probability distribution can be indicated by the corresponding cumulative probability value. These cumulative probabilities of the litho-layer thickness, the vertical resistance, and the hydraulic conductivity of litho-class EE-k are depicted in Figure 3.12. The same data for litho-class EE-kz are depicted in Figure 3.13. The data of these pictures are generated using the aquitard resistance before and after calibration of the groundwater flow model. If randomly drawn from a PDF, these cumulative probabilities are expected to be uniformly distributed. A large divergence from a uniform distribution may indicate that the a priori distribution does not coincide with the found improved data.

The uncertainties of the observations of the litho-layer thickness and the vertical

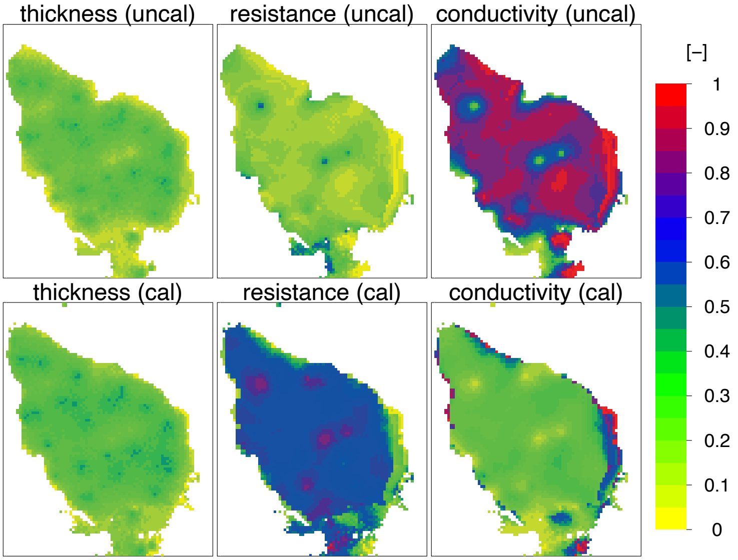

Figure 3.12: Cumulative probability of the most likely values of the uncalibrated (top) and calibrated (bottom) parameters of litho-class EE-k. The data is clipped at 0.05 m of the most likely thickness.

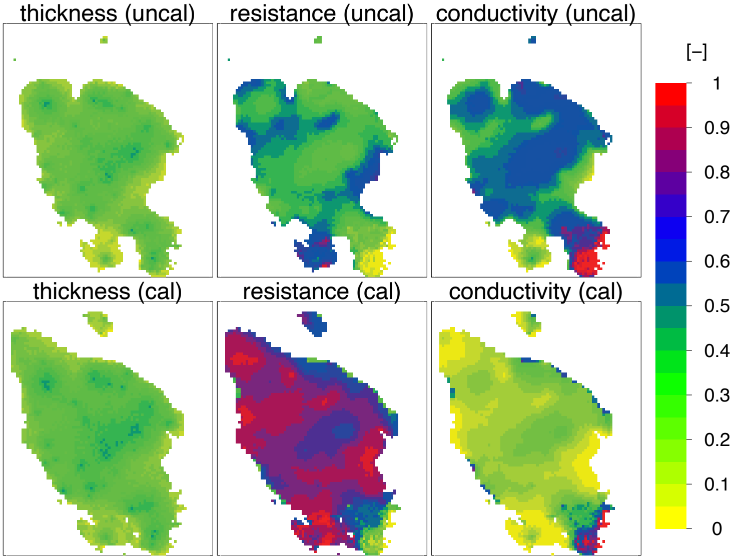

Figure 3.13: Cumulative probability of the most likely values of the uncalibrated (top) and calibrated (bottom) parameters of litho-class EE-kz. The data is clipped at 0.05 m of the most likely thickness.

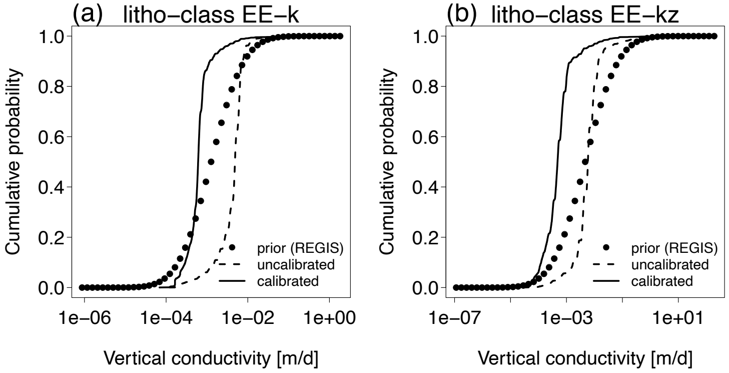

*Figure 3.14: Comparison of conductivity distributions of litho-class **(a)** EE-k and **(b)** EE-kz. Shown are the prior distribution of the REGIS system (dots), the distribution based on the uncalibrated C values (dashed line), and the distribution based on the calibrated C values (solid line). The x-axis is at log-scale.*

conductance are all represented by PDFs. In Figure 3.12, for litho-class EE-k, and in Figure 3.13, for litho-class EE-kz, it can be seen that the cumulative probabilities of the litho-layer thickness are mainly less than 0.5, which denotes the median of the PDF, for both the uncalibrated and the calibrated case. The maps of the vertical conductivities give a different picture. For litho-class EE-k, the majority of the values of the uncalibrated case are above 0.5, whereas the majority of the values for the calibrated case are below 0.5. The picture of the uncalibrated case of litho-class EE-kz (Figure 3.13) is less pronounced, only the lower right corner of the conductivity map shows some high values. This area may need some attention from the modelers. In the calibrated case the majority of the values is far below 0.5. So calibration reduces the conductivity and increases the thickness of the litho-layers, compared to the uncalibrated case. The distribution of the vertical conductivity is described in more detail in Section 3.3.2.

3.3.2 Distribution of improved conductivities

The a priori distributions of the litho-class conductivities, i.e. the distributions obtained from the REGIS information system, represent the best estimates given the available hydrogeological knowledge. One goal of the proposed method is to improve these distributions, or more specifically to decrease the uncertainty of the variables. The cumulative distribution functions (CDFs) of the conductivity values of litho-class EE-k and EE-kz, as discussed in the former section and depicted in Figures 3.12 and 3.13, are shown in Figure 3.14.

Herein, the a priori conductivity distribution of the REGIS system and the most likely conductivity distributions based on the uncalibrated and the calibrated resistance values are depicted. In fact, these distributions are spatial frequency distributions of the most likely values. Nevertheless, when applied to unobserved locations, these functions can act as a probability distribution. Hereafter, the distributions will be denoted by CDF or PDF.

In the study area, the majority of the calibrated resistance values is higher than the uncalibrated values. Consequently, the corresponding conductivity values must be lower or the layer thickness must be higher for the calibrated situation compared to the uncalibrated one. From Figure 3.14 it is clear that the conductivities from the calibrated case are much lower than the conductivities of the uncalibrated case. Only conductivity values with a corresponding most likely layer thickness greater than 0.05 m are used to create these CDFs. The presented conductivity distributions are derived at the scale of the used groundwater flow model. Since no full downscaling to core scale (borehole scale) is applied, these CDFs are valid at this model scale and can not be used as core scale distributions.

The results are based on a small study area and can currently not be extrapolated to the whole REGIS database.

3.3.3 Evaluation of a priori litho-layer thickness uncertainty

In this section, the selection of the litho-layer thickness uncertainty is justified. The observed litho-layer thicknesses of the available borehole data are expected to be uncertain, but the variance and distribution type are unknown. Nevertheless, the proposed method needs probability distributions of these observations. As stated in Section 3.1.2, we tested the effect of several a priori distributions to describe this thickness. The types of probability distributions tested are the Gaussian and log-normal distribution. Both distributions are tested with different values of the variance. The mean value of each distribution is set to the observed thickness.

For a given litho-layer the observations of the layer thickness fall into two groups: one group with the observed litho-classes and one group with the litho-classes observed absent. These groups are denoted as observed-thickness and zero-thickness, respectively. Two characteristics of the PDFs are important when judging the usability: the probability of negative values, and the width of the distribution. We defined the latter as the width of the 95 % probability interval, which is the distance between the 2.5 and the 97.5 % quantiles. In Table 3.4, an example is shown of the effect of the standard deviation assigned to the group of observed-thicknesses. The mean value presented is a round-off value as defined in Table 3.1. When applied to the litho-layers, the observed litho-layer thicknesses are used as mean value for the PDFs. Table 3.5 shows the same information for the zero-thickness observations.

Interpolation of the litho-layer thicknesses are performed using the settings as described above. The Gaussian distribution often yields negative most likely thicknesses (column percentile $< 0\,\mathrm{m}$) which makes this distribution unusable. Therefore, the log-normal distribution is used to describe the litho-layer uncertainty.

In Figure 3.15, the effect of the different variance settings on the interpolation of litho-class EE-kz, using log-normal distributions, is shown. In this figure, the maps in each row are calculated using the same zero-thickness variance. The maps in each column are calculated using the same observed-thickness variance. The maps

Table 3.4: *Effect of SD choise (low, medium high) of the observed-thicknesses distributions. The mean value of 1 m, which is the round-off value, is used as an example. Other round-off values show a proportional effect.*

SD	distribution	mean	s.d.	percentile	thickness	thickness	width
				$< 0\,\mathrm{m}$	2.5 %	97.5 %	95 %
		[m]	[m]		[m]	[m]	[m]
low	Gaussian	1.00	0.20	0 %	0.608	1.39	0.784
med	Gaussian	1.00	1.00	15.9 %	−0.961	2.96	3.92
high	Gaussian	1.00	5.00	42.1 %	−8.80	10.8	19.6
low	log-normal	1.00	0.20	0 %	0.665	1.45	0.781
med	log-normal	1.00	1.00	0 %	0.138	3.62	3.48
high	log-normal	1.00	5.00	0 %	0.0057	6.76	6.75

Table 3.5: *Effect of SD choice (low, medium high) of the zero-thickness distributions.*

SD	distribution	mean	s.d.	percentile	thickness	thickness	width
				$< 0\,\mathrm{m}$	2.5 %	97.5 %	95 %
		[m]	[m]		[m]	[m]	[m]
low	Gaussian	0.005	0.10	48.0 %	−0.190	0.201	0.392
med	Gaussian	0.050	1.00	48.0 %	−1.91	2.01	3.92
high	Gaussian	0.100	2.00	48.0 %	−3.82	4.02	7.84
low	log-normal	0.005	0.10	0 %	2.1×10^{-6}	0.0304	0.0304
med	log-normal	0.050	1.00	0 %	2.1×10^{-5}	0.304	0.304
high	log-normal	0.100	2.00	0 %	4.1×10^{-5}	0.608	0.608

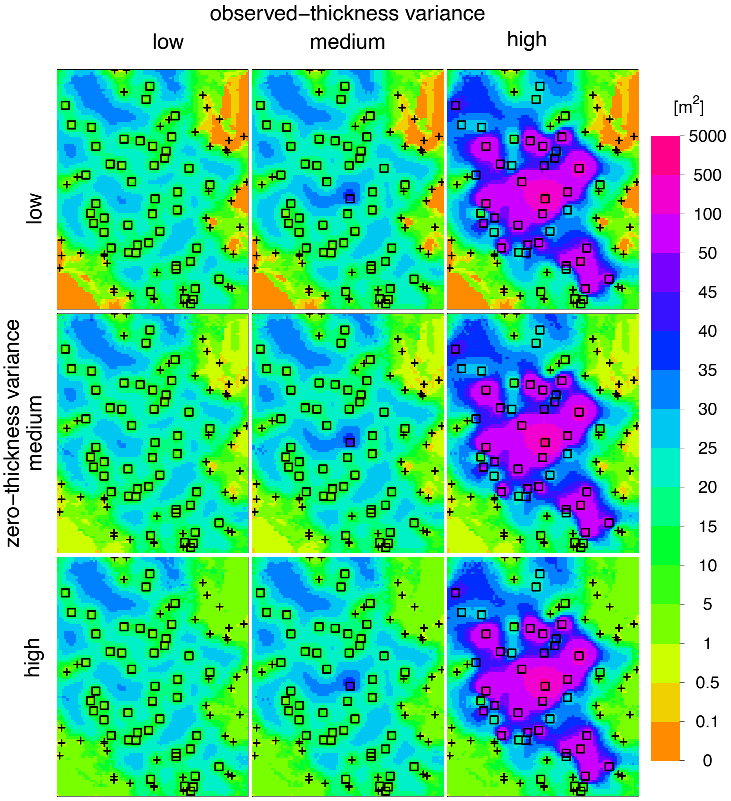

Figure 3.15: *Maps of interpolation variances of litho-class EE-kz for different settings of the observation variances. The used settings for the variances are shown in Tables 3.4 and 3.5. The squares denote the observed-thickness locations and the plus signs the zero-thickness locations.*

with high variance of observed-thickness (right column) show unlikely high variances at the area where this litho-class is present. Therefore, this variance setting is rejected. The difference between the low variance (left column) and the medium variance maps (middle column) is not very pronounced. The major difference is the sensitivity to observations with a high variance. In the center of the medium variance map (middle column), one observation location yields a very high variance. In one borehole, the litho-layer of litho-class EE-kz is here described with forty-nine sub-layers of one meter each, with each their own variance. After summation, this yields a very high variance for this location. For further processing the medium variance is used.

The lower-left and the upper-right corner of the study area are dominated by zero-thickness observations. The variances shown in these areas of the low (upper row) and medium variance (middle row) are low compared to areas dominated by the observed-thickness locations. Therefore, the high variance settings (lower row) for the zero-thickness observations are used for further analysis. Corollary, for further calculations the medium variance for the observed-thickness locations is used and the high variance for the zero-thickness locations.

3.3.4 Evaluation of kriging options

Depending on the nature of the observed data, and the associated assumptions of the underlying random field model, the appropriate form of data-transformation and kriging is chosen. Herewith, a decision has to be made whether or not to perform a data transformation. Hereafter, the decisions made are justified.

Layer thicknesses are, obviously, required to be greater than or equal to zero. Therefore, not every type of PDF is appropriate to describe the uncertain thickness. In the kriging interpolation with uncertain observations, two variables need to be assigned a probability density function: the observations of the layer thickness (Section 3.3.3), and the interpolation error. Usually, the interpolation error is assumed to be Gaussian distributed. No accurate information is available about the true shape of these PDFs. Therefore, the performance of the use of Gaussian and log-normal distributions was tested. Both distributions have their own deficiency, especially when the standard deviation is large compared to the mean value. In that case, the Gaussian distributions may yield negative thicknesses with too high probability, and the log-normal distribution may become very skewed. The latter is a disadvantage in finding representative most likely values because of the difference between the mode and the mean of the distribution. Because of the potential negative values of the Gaussian distribution, the log-normal distribution is tested for the interpolation error as well.

One way to avoid negative interpolated values is to transform the observations to their log values before interpolation, and back-transform them afterwards. Applied to block kriging, the different way the block average is calculated has to be considered. When kriging the log-transformed values, the block average is the geometric mean, kriging the non-transformed values yields the arithmetic mean. In

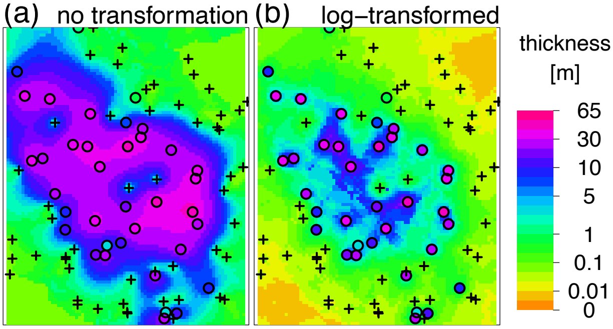

Figure 3.16: Mean value of interpolated layer thickness of litho-class EE-k. Shown is ***(a)*** *kriging without data transformation, and* ***(b)*** *kriging with log-transformed data. The circles denote observations with the litho-class present, the plus signs where it is absent. The circles are colored with the observed thickness.*

Figure 3.16, a comparison is made between interpolation of the thickness PDFs and the log-transformed thickness PDFs. With both methods, the interpolated thicknesses close to the observations are quite in agreement with the observed values. However, at larger distance the difference between the two methods is larger, with interpolated thicknesses from the log-transformed kriging being very low. From geological point of view this is not a feasible result. Even when the ranges of the variograms are increased, three times larger than derived from the data, the interpolated thickness remains much lower than presumed, given the observations. Thus the non-log-transformed kriging variant provides a better option.

3.4 Discussion

Since in the Netherlands the building of a hydrogeological model and its application in groundwater flow models is different from most other countries, the modeling context and the application of the proposed method is discussed here in more detail.

Usually, when building a groundwater flow model, a hydrogeological schematization dedicated to this groundwater flow model is defined simultaneously. Herewith, the building of the hydrogeological model is a part of the development of the groundwater flow model. In the Netherlands, the general purpose hydrogeological model REGIS is developed and maintained, independent of any particular groundwater flow model. REGIS covers the whole country up to a depth of a few hundred meters. REGIS is not built to serve as input for just one specific groundwater flow model, but it can be used in every study where hydrogeological subsurface information is involved. This is depicted in Figure 3.1. As mentioned in the introduction, the REGIS model is very detailed in the vertical direction and is therefore in most cases not suitable to serve directly as a hydrogeological schematization of a specific groundwater flow model. To serve the needs of a specific groundwater flow model a dedicated hydrogeological schematization is derived from REGIS. During calibration of the groundwater flow model this derived hydrogeological model is calibrated instead of the REGIS model. But the valuable information of the calibrated groundwater flow model should be incorporated in the REGIS model as well.

The proposed method aims to improve the hydrogeological model REGIS, by use of the improved information of a calibrated groundwater flow model. This implies that initially only the hydrogeological model REGIS benefits from this procedure and not the groundwater flow model. Moreover, the proposed method guarantees that the updated version of REGIS generates *exactly* the hydraulic parameters of the calibrated groundwater flow model. So this method has no application in an iterative calibration procedure of the groundwater flow model. Usually, a calibration contains an iterative procedure, but this is part of the calibration of the groundwater flow model itself. Nevertheless, the method can be used for detection of unlikely values during the calibration stage. Such unlikely values are an indication for errors in the hydrogeological model or elsewhere in the groundwater flow model.

In section 3.3.1 is shown that in some cases the most likely conductivity values, obtained from the calibration of the groundwater flow model, are not likely at all from (hydro)geological perspective. Whether or not a value is likely is decided by the modelers, but can be assisted by the proposed method. The unlikely results may indicate errors in the model identification. When this kind of results are found, discussions between the groundwater flow modeler and the (hydro)geologist may also lead to a modification of the REGIS model. Such an improved version of REGIS can be used to derive a new dedicated hydrogeological model for the groundwater flow model at hand. This feedback is indicated in Figure 3.1 by the dashed arrow from the most likely values to the REGIS model. In this way, the updating method described in this paper effectively becomes part of an iterative process of improved (hydro)geological schematization and groundwater modeling in which both geological and hydrological data are used. However, we stress that currently such an iterative procedure is not part of the REGIS work-flow. To make this happen we should also adjust the method to include errors in the transmissivities and hydraulic resistances resulting from the groundwater model calibration. This is part of ongoing research.

The incorrect presence or absence of, for instance, clay layers in the hydrogeological model can be corrected by the calibration through adjusting the vertical resistance. In fact this is an identification error of the model, but the proposed method is to some extent able to translate this new vertical resistance into a new layer thickness and conductivity value in the hydrogeological model.

Corollary, the calibrated groundwater flow model does not benefit from the reparameterization of the REGIS model, but future studies can make use of an improved version of the hydrogeological model.

3.5 Conclusions

This paper describes a method to improve the parameterization of a general purpose hydrogeological model REGIS in a formal way, using information of calibrated groundwater flow models. The parameterization of the aquifers and aquitards of these groundwater flow models are derived from the hydrogeological units of the general purpose hydrogeological model. Each hydrogeological unit consists of one or more litho-classes. The proposed method appears to be able to improve the estimates of litho-layers thickness and conductivity. From the uncertain hydrogeological data, described by a multidimensional probability density function (PDF), the most likely parameter values are derived given the information available from calibrated parameter values in groundwater flow models. The most likely values are the values at the mode of the multidimensional conditional PDFs. The proposed method is applied to layer thicknesses and vertical conductivities at litho-class support. Herewith, the most likely litho-layer thickness and vertical conductivity values are obtained for the studied aquitard.

In the REGIS database the a priori probability distributions of the vertical conductivity, for a given litho-class, are assumed location independent. However, this is not a limitation of the proposed method but of the available data. It is not unlikely that the a priori distribution of the vertical conductivity of a litho-class is spatially varying. When the spatial variability of the probability distributions is known, this should be used in the described method. Until know, when applied to a larger study area, this method can be used to find this spatial variability.

The spatial distributed most likely values of the vertical conductivity per litho-class are used to create an a posteriori distribution of the parameter. This distribution is compared with the corresponding a priori distribution. The a posteriori distributions of the two most important litho-classes show much less variability than the corresponding a priori distributions do. This reduction of uncertainty can be expected since additional knowledge is added using results from a calibrated groundwater flow model. Thereby, the obtained results hold for a small study area and for the modeling scale of the hydrogeological model and the groundwater flow models, whereas the a priori uncertainty information of REGIS is based on data of the whole data base and on a smaller scale. So, to be able to update the REGIS conductivity distributions a larger study area has to be processed and a downscaling method must be applied.

One way the results are evaluated is to show the position of the obtained most likely values in their a priori cumulative distribution function (CDF). This position is indicated by the corresponding cumulative probability. The most likely values of some parameters show a strong systematic deviation from the a priori distribution, with the majority of the values either lower or higher than the median of the a priori distribution. In case of a data update, the a posteriori distribution should of course divert from the a priori distribution, but a strong systematic deviation may indicate errors, either caused by data errors or a wrong perception about the hydrological system. The proposed method can thus serve as a tool to guide the discussion between experts from different domains.

With the described method, the most likely values are derived for each litho-layer separately, neglecting the thickness of adjacent layers. Obviously, it is not possible to change the thickness of a layer without affecting the thickness of the adjacent layers. The present study does not take this into account and only aims to describe a method to find the most likely combination of layer thickness and conductivity. A future study should account for all layers of the hydrogeological model, where the sum of all layer thicknesses is constrained.

It is quite likely that the litho-layer properties of horizontal adjacent grid cells are correlated. The presented method is applied at one grid cell at a time, so spatial correlation is not explicitly taken into account. Nevertheless, implicitly this correlation is present in the used data. The PDFs of the layer thicknesses are obtained by kriging interpolation, which has a smoothing effect. The used PDFs of the conductivities are everywhere the same within one litho-class. Herewith, the main source of lateral variability is the calibrated parameter field. The hydraulic response of the groundwater flow model, and the observations, may justify highly variable calibrated parameter fields. Since we stated that the calibrated values are considered to be the truth, variability of this parameter should be reflected in the results. In the results of our calculations, we haven't discovered parameter variability on short distance which could be caused by applying the method on single grid cells instead of taking the spatial correlation into account. Therefore, we think that neglecting the horizontal correlation is not of great importance to the results.

In the presented method, horizontal connectivity of litho-layers is neglected. Since litho-layers can have thicknesses at the order of centimeters, it is virtual impossible to find any connectivity at a distance between individual observations (boreholes). When only groundwater flow models are used in the feedback procedure, and no transport models, the horizontal connectivity of the litho-layers is of less importance, but especially in transport models it is an important issue. For now, this subject is left for future research.

The proposed method assigns a PDF to the thickness of every single litho-layer from the borehole descriptions. In Section 3.3.3 an example is shown where this yields an unlikely high variance for a thickness observation. When within one borehole adjacent litho-layers are of the same litho-class, aggregation of these litho-layers before assigning a variance may give a more appropriate representation of the uncertainty. This may yield a more realistic uncertainty description of the thickness observations, and an option for future application of this method.

As with the assignment of litho-classes, also the calibrated vertical resistance of the groundwater flow model is regarded as perfectly known. A valuable extension to the presented method is to account for uncertainty of the calibration results. If the calibration method is able to provide probability distributions of the calibrated parameters, these distributions can be used instead of the deterministic values in the presented method. Methods like Monte Carlo simulation can be applied to draw multiple values from the PDF. Herewith, multiple observations of the same property are generated, which can be used in the calculations. For a more direct solution additional research is needed. Furthermore, with the implementation of uncertain calibrated values, results from different calibrated groundwater flow models in the same area can be compared.

The use of piecewise linear PDFs, instead of parameterized PDFs, makes it possible to perform the necessary calculations without the burden of deriving intractable analytical solutions or resort to time-consuming Monte Carlo analysis. Herewith, many different calculations can be tested with relatively little effort.

4

Investigating lateral differentiation of hydrogeological parameter values in deposits

Abstract. Determination of the hydraulic properties of deposits of the subsurface is a recurring subject. It is even more challenging when sparsely or uncertain data are available. When hydrogeological models are constructed, uniform parameter values, such as hydraulic conductivity, are often assigned to deposits of a specific origin and age. It is, however, not likely that a specific deposit has the same hydraulic conductivity values everywhere. In the previous chapter, a method is developed to use calibrated subsurface parameters from a groundwater flow model to improve the parameterization of a hydrogeological model. In this chapter, that method is applied to an area in the Netherlands to investigate if a lateral differentiation in the hydrological parameterization of the same deposits can be found.

KNOWLEDGE about unconsolidated sediments in the subsurface is of great importance to (ground)water management. Usually, the knowledge about the presence and the properties of these sediments is only available at point scale (boreholes), or at small areas through derived data like ground penetrating radar [e.g. *Blindow et al.*, 2007] or seismic data [e.g. *Schuck and Lange*, 2007]. Exhaustive information of the sedimentary material of the entire area of interest, at the appropriate scale, is seldom available. To fill this gap, geological models are developed to serve as the basis to analyze, simulate and predict processes in the subsurface. In sedimentary basins, such models describe the geometry and properties of geological units. Such units are often described as layers with sediments of the same origin in time and space, and with similar lithological characteristics. Hereafter, these layers are called litho-layers, and the sediments are denoted by litho-class. Typically, the thickness of litho-layers described in geological models is in the order of centimeters to decimeters or meters. We focus on the processes of groundwater flow. Subsurface properties of particular importance for groundwater flow are the hydraulic properties such as hydraulic conductivity and storage capacity. A geological model dedicated to groundwater flow is called a hydrogeological model. An example of a hydrogeological model is the Dutch national hydrogeological model REGIS [*Vernes et al.*, 2005; *Vernes and van Doorn*, 2006]. The rational of having a national hydrogeological model is its multiple use. Therefore, it is important to continuously improve the hydrogeological model. In Chapter 3 a method is presented to find the most likely parameter values of a hydrogeological model, given the calibrated values of a groundwater flow model. That method is applied in this chapter.

The hydraulic properties in the hydrogeological model, such as the horizontal and vertical conductivity, are to a large extent based on core sample analysis, resulting in a probability density function (PDF) for each property and each litho-class. These distributions are a part of the REGIS model. Due to the limited number of data, this PDF is currently independent of the spatial coordinates. However, a litho-class can be present in an extensive area and spatial differences are likely from geological point of view. In a groundwater flow model the litho-layers are aggregated to form a limited number of aquifers and aquitards. Therefore, each aquifer and aquitard may consist of multiple litho-layers. Furthermore, to assign values to all grid cells of the groundwater flow model the layer properties are interpolated.

The values of the hydraulic parameters of the aquifers and aquitards in each grid cell are subject to uncertainty because of the interpolation, the PDF of the litho-classes, and the aggregation of the litho-layers. In order to increase confidence in the groundwater flow model, it is usually calibrated against observed groundwater heads. Basically, this means that the transmissivity of the aquifers and the resistance of aquitards are adjusted to arrive at an acceptable fit with the observed heads [e.g. *Zimmerman et al.*, 1998; *Valstar et al.*, 2004; *Carrera et al.*, 2005; *Hendricks Franssen et al.*, 2009; *Hoteit et al.*, 2012]. Since REGIS is a general purpose hydrogeological model, the calibration of a groundwater flow model has only effect on the derived hydrogeological model, and not on REGIS it originates from. In Chapter 3 we developed a feedback procedure to adjust the thickness and the conductivity values of the litho-layers in the general purpose hydrogeological model, given the calibrated value of the aquifer or aquitard of the groundwater flow model. This leads to different values of the conductivity of the same litho-class in different grid cells. In this chapter, we analyze the spatial pattern of the adjusted values of the conductivity of the litho-layers. We were able to identify spatially distinct subsets of a litho-class, with less variation of conductivity inside each subset, thus providing a better starting point for future use of the hydrogeological model.

In Section 4.1 the used models and the geological context are described. In Section 4.2 we first describe the methodology used. Next, in Section 4.3 we present an application of the method to an area in the Netherlands where a hydrogeological model as well as large scale groundwater model was available. Finally in Section 4.4 we discuss the results, and the potential applications and draw conclusions.

4.1 Material

4.1.1 Models

The Geological Survey of the Netherlands (TNO-GSN) develops and maintains a large information system with subsurface data and models. The models include a Digital Geological Model (DGM) [*Gunnink et al.*, 2013] which describes the geological units or formations, depth and extent of these units, based on a lithostratigraphical classification. The units are described up to a depth of about 500 m. Consistent with this DGM, the hydrogeological model REGIS [*Vernes et al.*, 2005; *Vernes and van Doorn*, 2006] is defined. The REGIS model describes the subsurface in terms of high and low conductivity model layers, the so called hydrogeological units. These units are based on the (assumed) hydraulic properties of the deposits. The presence or absence of the units is mainly based on the interpretation of the borehole descriptions and the classification of the units in the DGM. To all distinguished intervals in the borehole descriptions a litho-class is assigned. These litho-classes are defined by a combination of the geological formation and the lithological properties of the deposits, like sand, clay and peat. So a hydrogeological unit contains one or more litho-classes.

The hydrogeological model REGIS is designed to feed groundwater flow models

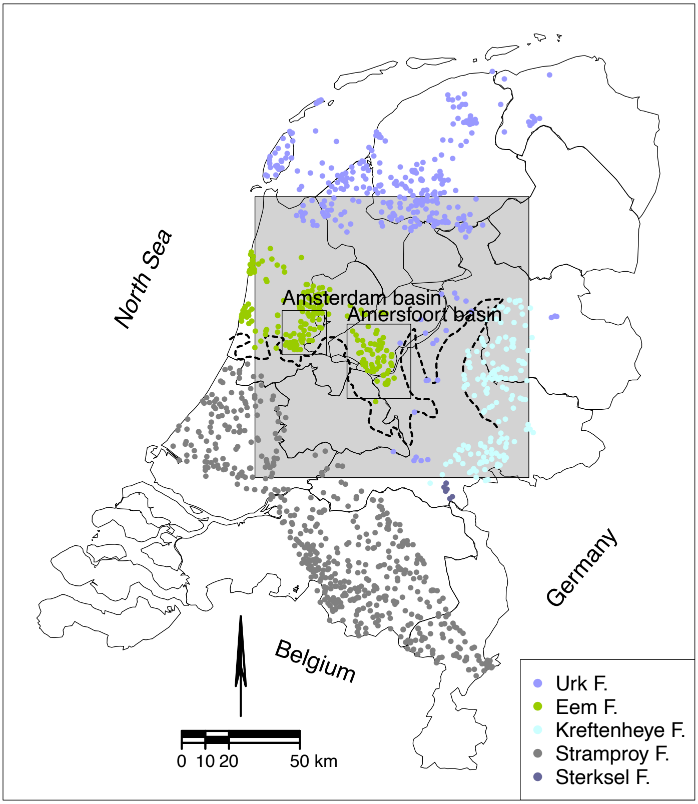

Figure 4.1: *Model area of the groundwater flow model (gray). The dots denote the boreholes in the fourth aquitard of groundwater flow model AZURE. The colors denote the Formation. The dashed line shows the approximate northern edge of the ice pushed ridge (after Peeters et al. [2016]) of the Saalian glaciation.*

with data for the subsurface description. For each new groundwater flow model, a dedicated hydrogeological model from the REGIS model is derived which meets the aim of the new model. Such a derived hydrogeological model contains typically up to about ten aquifers and aquitards, where REGIS defines over one hundred hydrogeological units. Therefore, multiple hydrogeological units are aggregated to form an aquifer or aquitard for each specific groundwater flow model. In case of calibration of the groundwater flow model, the derived hydrogeological model is calibrated and not the REGIS model. A method to conduct the feedback from the calibrated groundwater flow model to the REGIS model is extensively described in Chapter 3, and is briefly described in Section 4.2. This method is applied using the calibrated data of the AZURE groundwater flow model [*de Lange and Borren*, 2014] to find the most likely conductivities and layer thicknesses of the litho-classes for each grid cell of the REGIS model. The AZURE model is a groundwater flow model in the Netherlands. The extent of the model is depicted in Figure 4.1. This model is in the vertical direction discretized with nine aquifers and eight aquitards. The horizontal discretization is 100 m × 100 m.

In this study, the properties of one aquitard of the groundwater flow model are evaluated. This aquitard consist mainly of deposits of the last interglacial, the Eemian.

The models and data of DGM and REGIS are available through the DINO internet portal [*TNO-GSN*, 2021]. The data used in this study are not obtained through this portal but directly from the databases.

4.1.2 Geological setting

The model area of the groundwater flow model AZURE describes a large part of the Netherlands with multiple model layers. In this study, the proposed method is applied to one aquitard (aquitard 4) and only a part of the total groundwater flow model. In Figure 4.1, the AZURE model area is depicted by the gray shaded area, along with the boreholes in aquitard 4. The deposits in this aquitard are of different origin, which is depicted by the color of the dots.

The deposits of the Stramproy Formation originate from Belgium rivers and locally reworked deposits. In the north-west part of its extent also marine sediments are recognized as part of this formation. Deposition took place from the Early Pleistocene (Tiglien) until the lower Middle Pleistocene (Cromerien). The deposits consist predominantly of medium fine to medium coarse-grained sands, and less frequently of fine sands and clay or coarse sands [*Lang and Weerts*, 2003].

The deposits of the Sterksel Formation originate from the River Rhine and the Meuse. The deposits consists predominantly of moderate to coarse-grained sands, but also clay and fine sand are present [*Westerhoff*, 2003]. The majority of the deposits with low hydraulic conductivity are assigned to the third aquitard of the groundwater flow model. Only a few observations are assigned to the fourth aquitard and therewith of low importance to this aquitard.

The deposits of the Urk Formation originate from the River Rhine during the late Cromerien until the mid Saalian. The majority of this formation consists of medium-fine to very-coarse sands [*Bosch et al.*, 2003a]. The less frequent deposits, with lower hydraulic conductivities, are incorporated in the fourth aquitard of the groundwater flow model.

The deposits of the Kreftenheye Formation originate from the River Rhine during the late Saalian until the Early Holocene. The deposits consists predominantly of medium to very coarse-grained sands. Also fine sands, clay, and sporadically peat is found [*Busschers and Weerts*, 2003]. The boreholes which are used in the fourth aquitard, as denoted in Figure 4.1, belong to a subdivision of the Kreftenheye Formation, the Twello Member, with a prevailing lithology of fine to coarse sands and clay.

The sediments of the Eem Formation are of marine origin. Medium fine to very coarse sand are the prevailing lithology, but clay layers of tenth of meters are also present [*Bosch et al.*, 2003b]. The current study emphasizes on the deposits of the Eem formation. The majority of these deposits are found in the Central Depocentre

[*Peeters et al.*, 2015, 2016] which basin is formed during the Saalian glaciation. The south side of the basin is bounded by ice pushed ridges. This border is drawn as a dashed line in Figure 4.1. At the north-east side, the River Rhine entered the Central Depocentre [*de Gans et al.*, 2000; *Busschers et al.*, 2007], and deposited sediments in the northern part of the Central Depocentre [*Peeters et al.*, 2016]. The northern and southern part of the Depocentre are separated by a sill in the base of the Eemian sediments, this sill reaches a height of about 40 m below present sea-level [*Long et al.*, 2015]. During the Saalian glaciation two deep basins where formed in the Central Depocentre, the Amsterdam Basin and the Amersfoort Basin (Figure 4.1). The floors of both basins reach a depth of over one hundred meter below present sea-level [*Zagwijn*, 1983; *Cleveringa et al.*, 2000]. The basins were topographically separated from the sea by sills at a depth of about 35 to 40 m below present sea-level [*Zagwijn*, 1983; *Cleveringa et al.*, 2000; *de Gans et al.*, 2000; *Peeters et al.*, 2016]. During the late Saalian and the early Eemian, the deposition in the southern part of the Central Depocentre took mainly place under lake conditions [*de Gans et al.*, 2000]. After transgression, deposition in this area continued in a lagoonal environment [*de Gans et al.*, 2000; *Peeters et al.*, 2015]. A sill with a depth of about 25 m below present sea-level separates the two basins [*Zagwijn*, 1983], which influenced the infill of both basins. The pollen content of the sediments show that the marine deposition in the Amersfoort Basin starts later than in the Amsterdam Basin [*Cleveringa et al.*, 2000]. Also hiatuses in sediments are found in the Amersfoort locality boreholes, compared to the Amsterdam locality borehole [*Cleveringa et al.*, 2000; *Long et al.*, 2015], which suggests different sedimentation circumstances of the Eem Formation sediments in both basins. However, it can not be concluded what this exactly means for the sediment properties, like hydraulic conductivity, for the same litho-class in the different basins.

4.2 Methodology

In Chapter 3, a method is developed to find the most likely litho-class properties within each grid cell of a hydrogeological model. Hereafter, a functional description of this method is given.

A hydrogeological model can, and often does, serve as input data for a groundwater flow model. Such a hydrogeological model defines model layers, or hydrogeological units, of high and low hydraulic conductivity. In the Netherlands, a general purpose hydrogeological model REGIS [*Vernes et al.*, 2005; *Vernes and van Doorn*, 2006] is developed. This model serves as a hydrogeological model for multiple groundwater flow models. However, the number of hydrogeological units in REGIS, which is over one hundred, is usually too large to be workable in a groundwater flow model. Therefore, for each individual groundwater flow model a separate hydrogeological model is derived to meet the needs of this specific groundwater flow model. So multiple hydrogeological units are aggregated to form the aquifers and aquitards of the groundwater flow model. This collection of aquifers and aquitards is hereafter called the derived hydrogeological model. The parameterization of the (derived) hydrogeological model, is a first step in the modeling process of the groundwater flow model. Usually, the response of this initial version of the groundwater flow model does not replicate the groundwater observations accurately enough. Therefore, the groundwater flow model is calibrated, i.e. the parameterization of the derived hydrogeological model, to improve the quality of the subsurface parameterization, among other parameters. Much literature is available about the calibration of groundwater flow models, [e.g. *Zimmerman et al.*, 1998; *Valstar et al.*, 2004; *Carrera et al.*, 2005; *Hendricks Franssen et al.*, 2009; *Hoteit et al.*, 2012], but in this study no calibration of the groundwater flow model takes place. Instead a readily calibrated groundwater flow model is used. With a known connection between the derived hydrogeological model and the REGIS model layers, the calibrated values can be used to improve the property values, layer thickness (D) and hydraulic conductivity (K), of each litho-class of the hydrogeological model. So the knowledge added by the calibration of the groundwater flow model can help to improve the quality of the hydrogeological model REGIS. An important requirement is that the connection between the model layers of the hydrogeological model and the aquifers and aquitards is known.

The applied method starts at the borehole descriptions in the study area. To each interval of all borehole descriptions a litho-class identification is assigned. Each interval is also assigned to a specific hydrogeological unit. Herewith, the connection between all described intervals in the borehole descriptions and the corresponding aquifer or aquitard of the groundwater flow model is known. In the presented study, only the properties of an aquitard are investigated. Therefore, this method description will emphasize on the aquitard properties vertical hydraulic conductivity (K), layer thickness (D), and vertical resistance ($C = D/K$). Nevertheless, the method is applicable to aquifers as well. In most deposits, the vertical hydraulic conductivity (K_v) is different from the horizontal conductivity (K_h). Since in this study only the properties of an aquitard are discussed, we do not make this distinction here and only use variable K.

For each litho-class at each borehole location, the total layer thickness within the processed aquitard is determined. Usually, an aquitard consists of multiple litho-classes. These thicknesses are the observations at point scale. Subsequently, the point scale values are interpolated to grid cell average thicknesses. Along with the interpolation, the uncertainty of this interpolation is assessed. This uncertainty is described by a probability density function (PDF). To calculate the vertical resistance of a layer, the vertical conductivity is needed as well. Since this value is also uncertain, it is also described by a PDF. The PDF of the vertical conductance of each litho-class is obtained from the REGIS information system [*Vernes et al.*, 2005; *Vernes and van Doorn*, 2006]. In this information system, for each litho-class such a PDF is defined.

Within one grid cell, the total vertical resistance of an aquitard can be calculated as

$$C = \sum_{i=1}^{n} C_i = \sum_{i=1}^{n} D_i/K_i, \tag{4.1}$$

where n is the number of litho-classes within the aquitard. The variables C_i, D_i and K_i are the vertical resistance, the layer thickness and the vertical hydraulic conductivity of litho-class i, respectively. These variables are all defined as random variables (RVs). All variables D_i and K_i form an $2n$-dimensional joint distribution of C. In this joint distribution, the combination of values of D_i and K_i with the highest probability density, i.e. the mode of the distribution, can be found. If the vertical resistance c_m, as obtained by the calibration of the groundwater flow model, is assumed to be the true value of C, the mode of the joint distribution can be searched for, conditional to $C = c_\mathrm{m}$. With finding this mode, the most likely values of the marginal variable D_i and K_i are found.

In this study, this method is applied to an aquitard as defined in the AZURE groundwater flow model [*de Lange and Borren*, 2014]. After finding the most likely parameter values for each grid cell for each litho-class, the parameter values of the hydraulic conductivity K are drawn on a map. From these maps, for some litho-classes, areas with different ranges of conductivity can be distinguished. The differences in hydraulic conductivity can indicate differences in the sediments. Although deposits of the same litho-class are considered to have the same properties, lateral variation may become distinct with this method. In the next section the results are discussed.

4.3 Results

The fourth aquitard of the AZURE groundwater flow model is processed to obtain the most likely parameter values. These parameters include the most likely vertical hydraulic resistance (C) for each litho-class and, subsequently, the most likely layer thickness (D) and hydraulic conductivity (K). It is of particular interest or lateral differences in litho-class properties are present or not.

4.3.1 Vertical conductivity

Each litho-class is a combination of a geological formation and the lithology (clay, sand, peat, etc.), in total 36 litho-classes with a more or less significant contribution to aquitard 4 were found in the borehole data. Due to different depositional processes, different separated groups (formations) of litho-classes are present at sub-areas, as shown in Figure 4.1.

Figure 4.2 shows the vertical hydraulic resistance of aquitard 4 as used in this study. This hydraulic resistance is based on the parameterization of the REGIS hydrogeological model and the subsequent calibration of the groundwater flow model AZURE. The resistance of 1 day is assigned to the grid cells where the aquitard is absent. In the applied method in this study, this calibrated vertical resistance is assumed to be the true value for this aquitard. As can be seen, the aquitard does

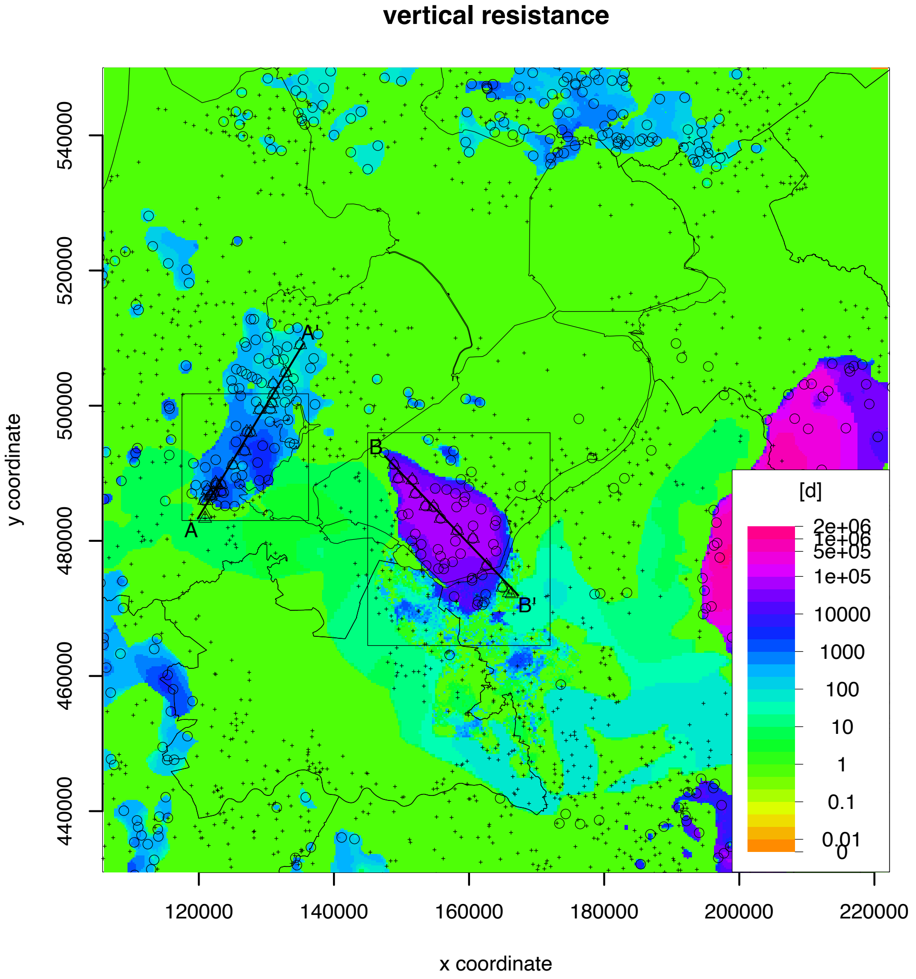

Figure 4.2: *Calibrated vertical hydraulic resistance of aquitard 4 of the AZURE groundwater flow model.*

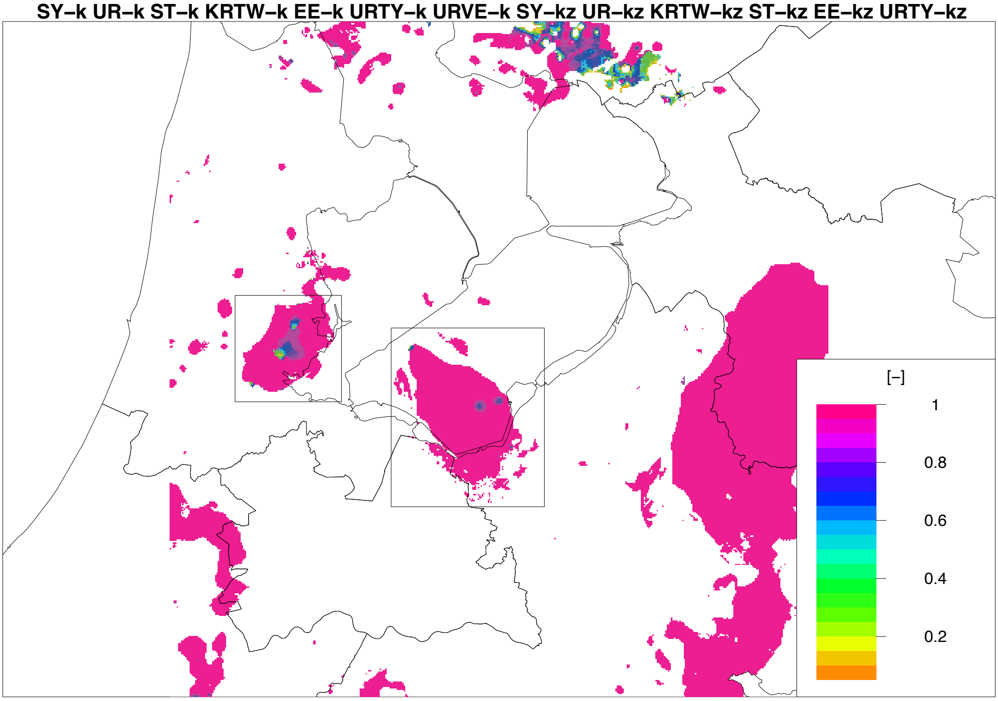
SY–k UR–k ST–k KRTW–k EE–k URTY–k URVE–k SY–kz UR–kz KRTW–kz ST–kz EE–kz URTY–kz
[–]
1
0.8
0.6
0.4
0.2

Figure 4.3: Fraction of total resistance explained from the most likely resistance values of the clay and sandy clay lithoclasses in the aquitard. Only data with corresponding layer thickness greater than 0.01 m, and resistance greater than 1 day are presented. The header of this picture shows the included litho-classes.

not have a high vertical resistance everywhere in the model area. Although these deposits of different formations are treated as one aquitard in the groundwater flow model, the deposits of different formations do not overlap and can be interpreted separately.

Seven classes of different lithology are recognized, i.e. clay (k), sandy-clay (kz), fine sand (zf), mean sand (zm), coarse sand (zg), peat (v) and gravel (g). These abbreviations are taken from the Dutch terminology, but are maintained to stay consistent with the REGIS database notations. Although an aquitard is considered, there are some litho-classes that consist of coarser material (mean sand to gravel) with atypical hydraulic properties for aquitards. Since these litho-classes are embedded in layers of higher hydraulic resistance, they are modeled as part of the aquitard. Nevertheless, the litho-classes with lithology mean sand, coarse sand and gravel do not contribute significantly to the vertical resistance. The most likely vertical resistance of these three classes, after application of the proposed method, do have a contribution of less than one percent of the total vertical resistance of the aquitard. Figure 4.3 shows the sum of the most likely vertical hydraulic resistance of the clay and sandy-clay lithologies as a fraction of the total aquitard resistance. It should be noted that the data of the results is only shown for grid cells where

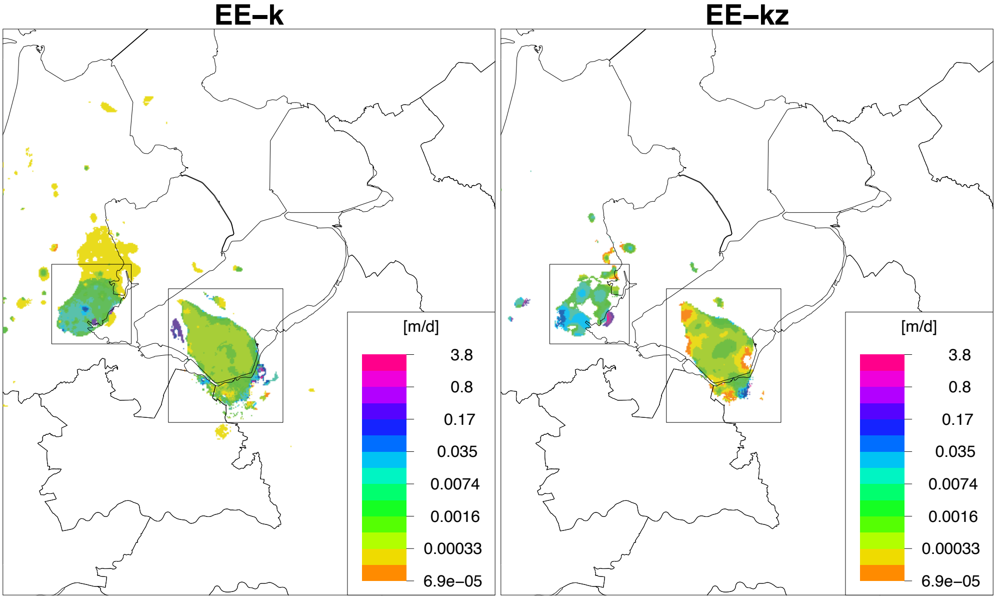

Figure 4.4: Most likely vertical hydraulic conductivity of the clay deposits of the Eem Formation, litho-class clay (left) and sandy clay (right).

the total vertical hydraulic resistance is greater than 1 day, and the total most likely layer thickness is more than 0.05 m, and the thickness of the individual litho-layers should be more than 0.01 m. As can be seen in Figure 4.3, almost all resistance can be explained by these two lithologies. At locations with a fractions lower than 1, other lithologies have a more or less significant contribution to the aquitard resistance. The missing part is almost completely explained by the peat lithology. The sand and gravel fractions do not have any significant contribution to the hydraulic resistance.

In Figure 4.4 the most likely vertical hydraulic conductivity of the clay and sandy-clay lithologies of the Eem Formation are depicted. The two larger separated areas show a different range of conductivity values in both litho-classes. These areas happen to coincide with the Amsterdam Basin and the Amersfoort Basin (Figure 4.1). For the uncalibrated hydraulic resistance of aquitard 4, this difference in hydraulic conductivity does not appear in the data. So this lateral difference in most likely parameter values is caused by the calibration results, and this shows how, through the groundwater model, head observations can inform on the spatial variability of conductivities within the same formation. The most likely conductivity values of both basins are depicted as a cumulative frequency distribution in Figure 4.5 for both, the calibrated and the uncalibrated data. It is clear that for the uncalibrated case the distributions do not differ much, but for the calibrated case that the conductivity in the Amersfoort Basin is much lower than in the Amsterdam Basin.

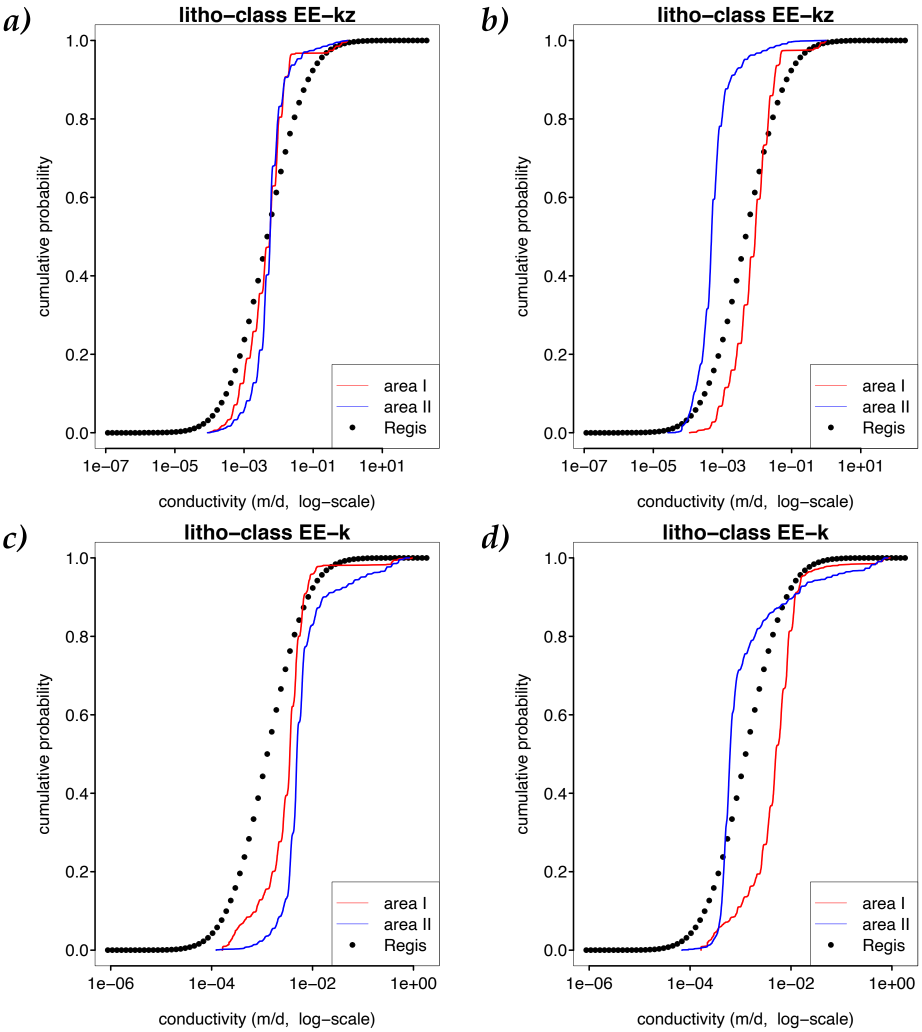

Figure 4.5: *Frequency distributions of the most likely conductivity values of area I (Amsterdam Basin) and II (Amersfoort Basin). Pane a): litho-class EE-kz, uncalibrated data, b): litho-class EE-kz, calibrated data. c): litho-class EE-k, uncalibrated data, d): litho-class EE-k, calibrated data. The CDFs of REGIS (black dots) are added to the graph for comparison, these graphs show the uncertainty of the conductivity values of a litho-class. In REGIS the mean values of these distributions are used for parameterization.*

4.3.2 Layer thickness

Besides the hydraulic conductivity, the most likely layer thickness is obtained. Figure 4.6 shows the most likely total layer thickness of all litho-classes. This map also shows the observations with the total layer thickness (circles), and the locations where the aquitard is absent in the borehole descriptions (+ signs). In the cross sections in Figures 4.7 and 4.8 for each litho-class the most likely layer thickness is depicted. Boreholes nearby the cross section are projected on the cross section. The depicted layer thickness in a borehole for a litho-class is the total layer thickness in that borehole for that specific litho-class. Which means that the layer thicknesses of all layers with the same litho-class are summarized to one thickness. So the order of the litho-classes and the number of described layers is not shown within a borehole. The black dashed line in the cross sections shows the layer thickness as modeled in the REGIS hydrogeological model, and as is applied in the AZURE groundwater flow model. Often, the most likely layer thickness is less than the REGIS layer thickness, which can be explained by the proximity of zero-thickness boreholes (Figure 4.6, + signs) during the interpolation.

4.4 Discussion and conclusions

A method is applied to identify lateral differences in hydraulic conductivity within deposits of a certain lithology within the the same geological formation (litho-class). In this study, an aquitard containing deposits of several litho-classes is processed. The total vertical hydraulic resistance of the aquitard, which is an input variable of the method, is obtained from a calibrated groundwater flow model. The applied method searches for the most likely layer thickness and hydraulic conductivity of deposits of different lithology (litho-class) within an aquitard.

Two litho-classes of the Eem Formation, the clay and the sandy-clay litho-class, show a difference in the most likely vertical hydraulic conductivity in two distinct areas. These two areas coincide with two depositional basins of the Eem Formation, i.e. the Amsterdam Basin and the Amersfoort Basin. From the literature it is known that the infill of these basins took place under slightly different circumstances and timing. Both basins were during the infill (partly) separated by sills, which caused a different sedimentation climate. Comparisons in the literature of the two basins show that certain layers of deposits are present in the Amsterdam Basin but are missing in the Amersfoort Basin. So it is likely that these differences in the two basins cause the differences in most likely hydraulic properties of the two basins. Nevertheless, a convincing geological explanation for the differences in hydraulic conductivity of the Eem Formation deposits at the different locations can not be claimed yet.

The method is also applied using the uncalibrated vertical hydraulic resistance as input. From this data, no difference in most likely hydraulic conductivity between the two basins is found. So through the groundwater heads the calibration of the groundwater flow model adds information to the data which exposes the

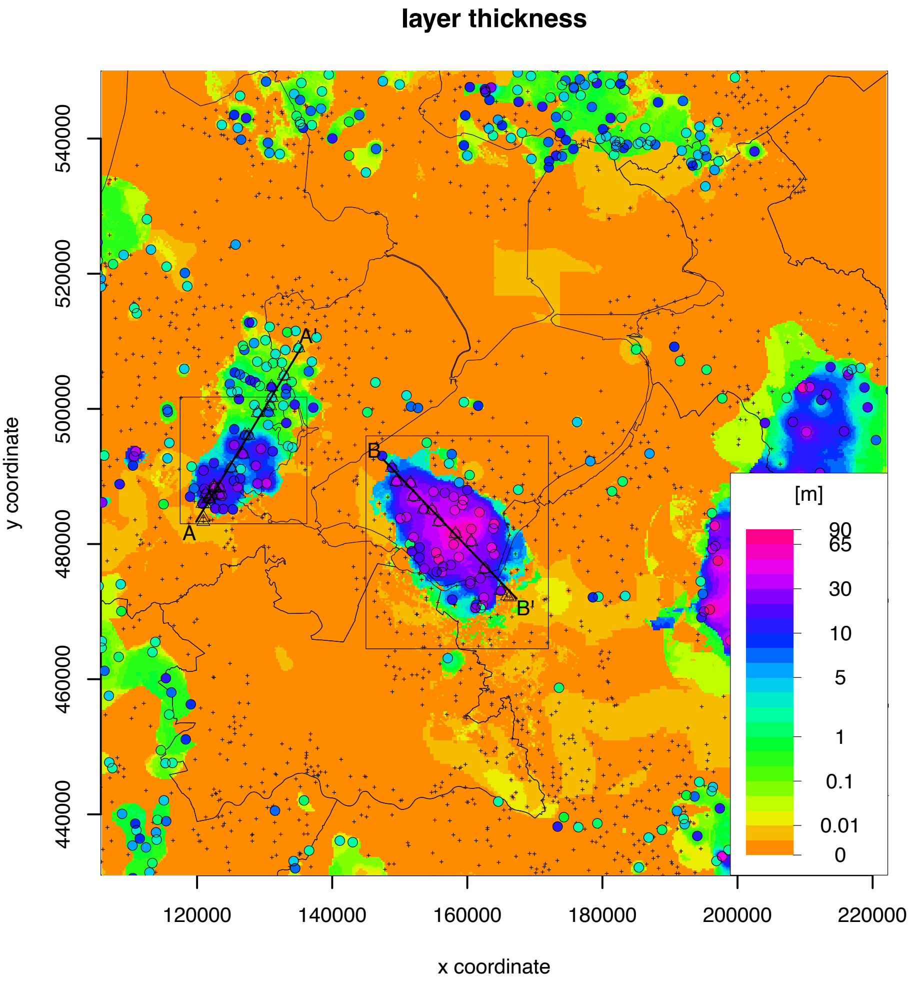

Figure 4.6: Map with total most likely layer thickness and the location of the cross sections. The circles show the total layer thickness at the borehole locations. The plus signs denote the locations where aquitard 4 is observed absent. The triangles denote the locations which are projected in the cross sections.

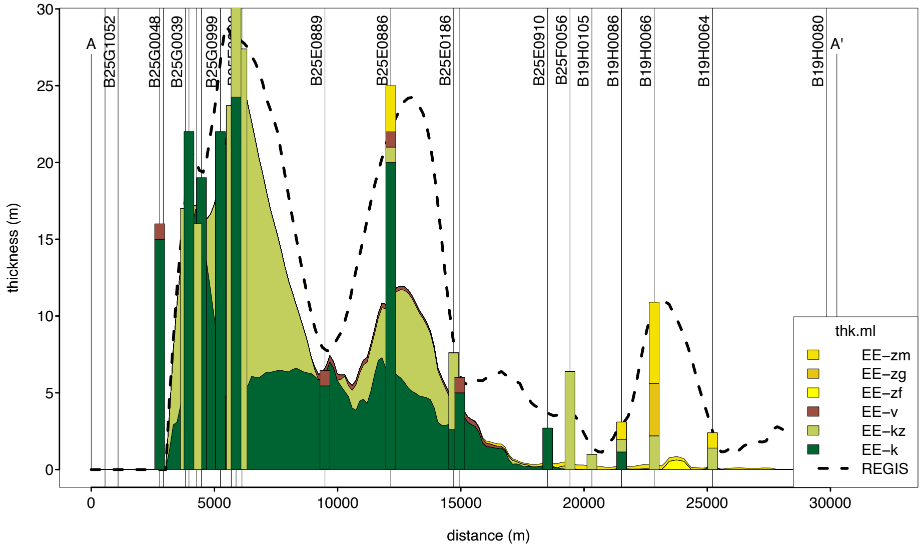

Figure 4.7: *Cross section with the cumulative most likely layer thickness for each litho-class through the Amsterdam Basin (A-A′). The dashed line shows the corresponding layer thickness of the REGIS-II hydrogeological model.*

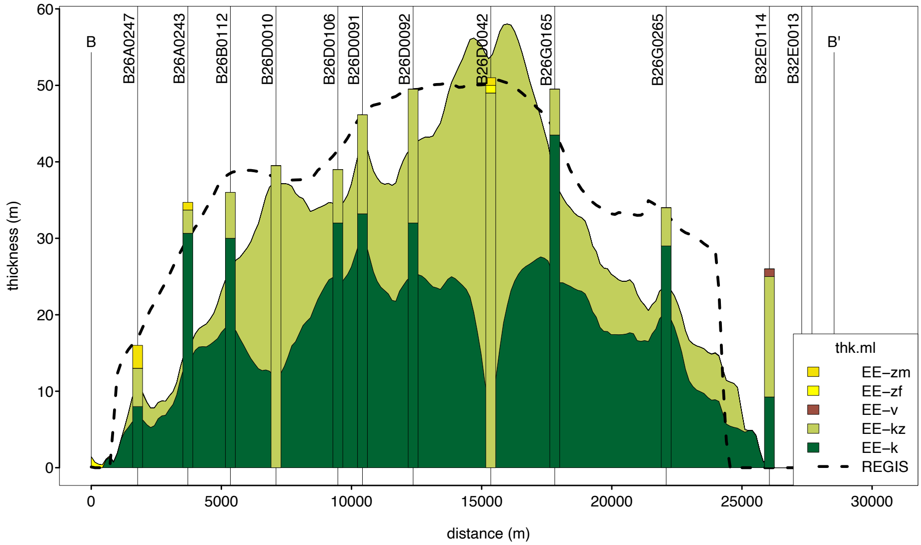

Figure 4.8: *Cross section with the cumulative most likely layer thickness for each litho-class through the Amersfoort Basin (B-B′). The dashed line shows the corresponding layer thickness of the REGIS-II hydrogeological model.*

differences in the hydraulic properties.

The most likely layer thickness for each litho-class is determined too. It appears that the total most likely layer thickness is in most locations less than the layer thickness as presented by the REGIS hydrogeological model and that has been used in the AZURE model. A suggestion may be to adapt the method with the possibility to use layer thickness of the aquitard, obtained from a independent source, as an extra constraint.

5

Improving a hydrogeological model with a Bayesian approach

Feedback with full uncertainty

Abstract. In this chapter, a method is presented to use the results of a calibrated groundwater flow model to estimate updated probability density functions of layer thicknesses and hydraulic conductivities of the underlying hydrogeological model. To this end, a Bayesian network of aquitard properties, specifically layer thickness and hydraulic conductivity, is built and evaluated. Such an aquitard is composed of multiple sub-layers with each their own parameterization. Since the calculation of the aquitard hydraulic resistance is highly deterministic, the Bayesian network mainly consists of deterministic nodes. Probability density functions of the layer thickness and the hydraulic conductivity are available from previous studies, and these distributions act as prior knowledge of the network. The aim is to update these prior distributions using observations of the total aquitard resistance. As observations calibrated aquitard resistances of a groundwater flow model are used, which can also be subject to uncertainty, allowing the inclusion of the results of multiple calibrated groundwater models. All probability density functions are described by piecewise linear functions, which makes the evaluation of the network independent of the shape of the distributions.

HYDROGEOLOGICAL MODELS describe properties of the subsurface in a geological and hydrological context, and are developed to serve as a base for further research and application. An important application of hydrogeological models is the hydraulic parameterization of the subsurface for groundwater flow models. Like all models, hydrogeological models never describe reality exactly but have, beside conceptual errors, a parameterization which is to some extent subject to uncertainty. So, when applied in groundwater flow models, the hydrogeological parameterization is usually calibrated to decrease the uncertainty and correct errors. In general, calibration of a groundwater flow model implies improving the hydrogeological model. In The Netherlands, this situation is different. The nation wide hydrogeological model REGIS [*Vernes et al.*, 2005; *Vernes and van Doorn*, 2006] is developed to serve as a general purpose model. From this REGIS model, dedicated models are extracted for specific applications, like groundwater flow models. After calibration, the extracted models are improved, but the REGIS model is unchanged. In Chapter 3, a method is described which finds the most likely parameter values (layer thickness and hydraulic conductivity) for the REGIS model, given a calibrated groundwater flow model. Herewith, an improvement of the parameterization of the calibrated groundwater flow model can be used to improve the parameterization of the REGIS model.

A shortcoming of the method of Chapter 3 is that only one source of calibrated data can be used to improve the REGIS model, which consequently results in only one most likely parameter value of the REGIS model. This one value is a good approximation of the update of the REGIS values, but it does not give information about the uncertainty of this most likely value. Also, in this method it is assumed that the calibrated values represent the true parameter values. Obviously, also the calibrated values are subject to uncertainty and it would be better to quantify this uncertainty and use it in the feedback procedure as well. In addition, in the same area multiple calibrated models may be developed and it would be beneficial if the results of multiple calibrated groundwater models could be used to improve parameterization of REGIS. In this chapter, a method is developed which honors all these available data and results in full probability density functions of updated hydrogeological parameters. To this end, a feedback procedure in a Bayesian context is proposed which honors prior and posterior uncertainty of the model parameters. In addition, the application of uncertain observations in a Bayesian Network is described.

In the literature, probability theory is widely used and described. In different areas of research a variety of notations and styles are used, but they often describe the same phenomena. In this chapter, applications and ideas are drawn from the variety of literature types. To see the connection between the different styles, an overview is given in Appendix C.

The REGIS model has a relatively high vertical resolution, where over one hundred hydrogeological units are defined. This resolution is usually too high to be used in groundwater flow models. Moreover, since the REGIS model is a nation wide model, not all defined hydrogeological units are present in all areas. Therefore, multiple hydrogeological units of REGIS have to be aggregated to one layer (aquifer or aquitard) to be applicable in a groundwater flow model. A hydrogeological unit may consist of multiple lithologies, like sandy clay, fine sand or peat. Layers of these lithologies of a certain formation or member (part of a formation) are hereafter called litho-layers. The calibration of the groundwater flow models is performed at the level of aquifers and aquitards. So for the feedback procedure, one calibrated value of the groundwater flow model represents multiple hydrogeological units of the REGIS hydrogeological model. Since all parameters (layer thickness and hydraulic conductivity) of the hydrogeological units are described by probability density functions (Chapter 3), this feedback problem can be adequately described using hierarchical Bayesian models or Bayesian networks [e.g. *Gelman et al.*, 2014, Ch. 5; *Bishop*, 2006, Ch. 8]. Such Bayesian models find wide application in various fields of study. In this chapter, Bayesian models are applied in the feed back procedure of the calibrated data to the hydrogeological model REGIS.

Although Bayesian models may seem straight forward, their application to the above proposed feed back procedure is far from straightforward. Firstly, we have probability density functions (PDFs) of the litho-layer thicknesses and the hydraulic conductivities, but these PDFs don't represent stochastic models in the Bayesian context. Therefore, we treat these PDFs as prior predictive distributions and decompose these into full stochastic models. This decomposition is performed by defining a location-scale parameter model with suitable prior distributions. Secondly, the calculation of litho-layer vertical resistance (quotient of the layer thickness and the conductivity) and the calculation of the total aquitard vertical resistance (sum of the layer resistances) are deterministic operations from a Bayesian point of view [*Cobb and Shenoy*, 2005, 2006; *Cinicioglu and Shenoy*, 2009], nevertheless these variables represent random variables. Because of the deterministic variables, a joint distribution of all parameters can not be formulated. Since we only have observations of the aquitard resistance, these can not directly be applied to the stochastic model. In the Bayesian belief literature, methods are described to solve this problem [*Cobb and Shenoy*, 2006; *Cinicioglu and Shenoy*, 2009; *Shenoy and West*, 2011; *Cobb and Shenoy*, 2017]. In this chapter we use and adapt such methods to make use of the observations of deterministic variables. Thirdly, observations are to some extent uncertain but are usually treated as exact. In this chapter, we describe a method to use uncertain observations in the inference procedure. Finally, calculation of the posterior distributions of the Bayesian model includes marginalization of the joint distribution function. Only a limited number of distributions do have a closed form solution for this marginalization. Performing marginalization for other distributions need to reside to numerical solutions of the integrals. In the literature, approximations of the distributions with exponential functions [*Rumí and Salmerón*, 2007; *Langseth et al.*, 2009] or polynomials [*Shenoy and West*, 2011; *Cobb and Shenoy*, 2017] or discretization of the continuous functions [*Kozlov and Koller*, 1997; *Neil et al.*, 2012] are applied. As in the previous chapters, here the continuous distributions are approximated by piecewise linear functions. The application of piecewise linear PDFs makes the calculations more tractable without the need of Monte Carlo simulations.

5.1 Methodology

In this section, a method is described to update the prior information of an aquitard of a hydrogeological model (in our case REGIS), given additional information. Usually, a hydrogeological model contains aquitards and aquifers, but for conciseness the method is only explained with an aquitard as an example. Nevertheless, the method applies for aquifers as well. Only the probability density functions (PDFs) of the hydraulic conductivity and the layer thickness of each litho-class are used to define the hydrogeological model, which together make up the aquitard vertical hydraulic resistance, and an observation of the total vertical resistance. The problem is approached by converting these data into a Bayesian model, and update the prior distributions using additional data.

In this chapter, all PDFs are denoted by the function $f(\cdot)$ without any subscript. When it is not clear from the arguments which variable the function describes, a subscript will be added.

5.1.1 Overview

The described methodology in this chapter is split up into a few steps. As a first step (Section 5.1.2), a Bayesian model (Figure 5.1 step 1, Figure 5.2) is defined which describes the dependencies between the hydraulic conductivities and the layer thicknesses on the one hand and the total aquitard resistance on the other hand.

The probability distributions of conductivities and layer thicknesses are known, but they need to be turned into a stochastic model by defining their respective marginal distributions (Figure 5.1 step 2). Since no parameterized PDFs are used, a method to perform this task is developed in Section 5.1.3.

A Bayesian graph represents a joint probability distribution and is defined for making inferences given evidence or observations. This joint distribution and the method of inference is defined in Section 5.1.4 (Figure 5.3). The Bayesian graph contains deterministic variables, and all distributions are described by piecewise

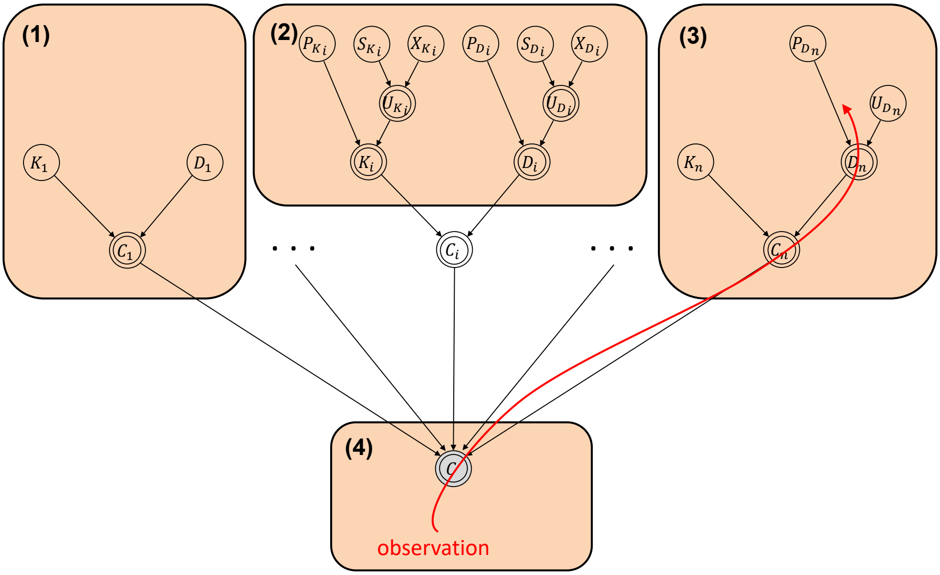

Figure 5.1: *Overview of the main steps in the methodology. Step (1) represents the setup of the Bayesian Network (Section 5.1.2). Step (2) shows the expansion of the marginal variables of conductivity (K) and layer thickness (D) into a stochastic model (Section 5.1.3). Step (3) shows the marginalization of the nuisance parameters in favor of the update of marginal variable P_{D_n} (Section 5.1.4). Step (4), the red arrow shows the path of the information of an observation to obtain the likelihood function (Section 5.1.5), and, in this example, the posterior of P_{D_n}.*

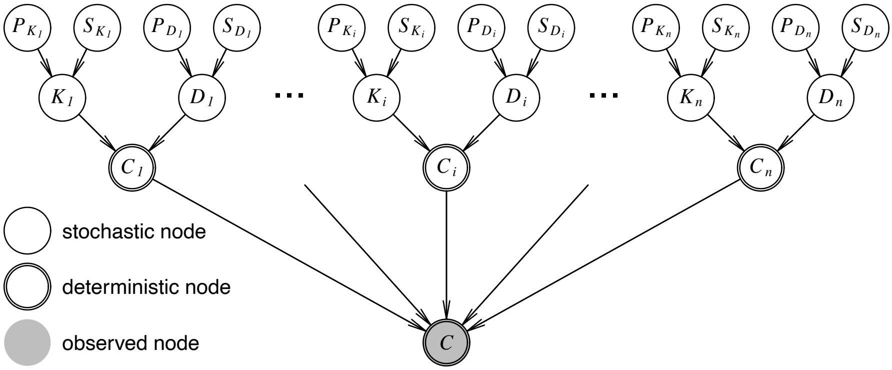

Figure 5.2: *Bayesian graph of the described problem. The nodes with a single border are stochastic nodes, the nodes with a double border are deterministic nodes. The gray shaded node contains the observed variable.*

linear PDFs. This calls for methods to handle these kinds of variables. Therefore, the implementation of deterministic variables in a joint distribution, the marginalization (Figure 5.1 step 3, Figure 5.4) of nuisance variables (variables temporarily beyond interest), and the definition of the likelihood functions are described.

Since the likelihood functions are also described by piecewise linear functions, an adequate discretization of these functions is important. The derivation of thereof is described in Section 5.1.5.

Usually, observations in a Bayesian inference are used as if they are deterministic. In case of multiple observations it may be beneficial to weigh them differently. In Section 5.1.6, a method is described to calculate the likelihood function with observations which are defined as random variables (Figure 5.1 step 4).

5.1.2 Bayesian network

A Bayesian network or Bayesian graph is a useful common graphical tool to present a probability problem, and a structured way to solve an inference problem. Such a network is drawn as a directed acyclic graph (DAG) [*Pearl*, 1986; *Bishop*, 2006, p. 362], where the relations between nodes are denoted by edges (lines). Each node represents a random variable (RV), and each edge the relation, or conditional dependency, between two RVs. In this context, directed and acyclic means that a strict hierarchical ordering exists between the nodes, which ordering is denoted by arcs (arrows) between the nodes. Hence, a path from any node in the network, following the arcs, will never end in the same node. This ordering of the nodes is a typical feature of a Bayesian network, in contrast to other graphs.

The inference problem in this chapter is finding the posterior distributions of the hydraulic conductivity and layer thickness of litho-layers, given an observation of the total vertical resistance of an aquitard. In Figure 5.2 this problem is depicted as a Bayesian network. In this graph, all nodes are depicted as circles, and the dependencies between the nodes by arcs. Herein, a parent of a node is the node with an arrow pointing to that node. For instance, in Figure 5.2 the RVs P_{K_i} (location parameter or mean) and S_{K_i} (scale parameter or standard deviation) are the parents of RV K_i (hydraulic conductivity). Similarly, a child node is a node which is pointed to by another node. Thus K_i is a child node of P_{K_i} and S_{K_i}. This network in Figure 5.2 consists of several types of nodes which all represent a random variable with an accompanying probability distribution. The single bordered circles define stochastic nodes, and the double bordered circles deterministic nodes. The definition of a deterministic node or variable is that it is completely defined by its parents without any additional uncertainty [*Shachter*, 1988; *Cobb and Shenoy*, 2005, 2006; *Cinicioglu and Shenoy*, 2009; *Cobb and Shenoy*, 2017]. Nevertheless, it is still a random variable. An example of a deterministic variable is the sum of two RVs (the parents). This sum yields a deterministic RV with a positive variance. In other words, if the values of the parents are known, the sum is also exactly known and has zero variance. From this definition, we may conclude that a node without parents (a so called leaf node) always is stochastic, as long as it is not a degenerate random variable (an RV with zero variance). Consequently, when integrating out a leaf node into a new leaf node, the result will always be a stochastic node. In the literature, the difference between a stochastic and deterministic variable is quite strict. In the next section (Section 5.1.4), we show that, when making some knowledge explicit, a stochastic variable may easily be converted into a deterministic variable.

In Figure 5.2, the node of variable C is shaded, which means that this variable is observed. This node has no descendants (no child nodes) and is usually called the root node. The main goal of the Bayesian inference is to propagate the information of these observations through the network to update the distributions of the leaf nodes. In the current study, no direct observations of variable C are available, instead calibrated values of a groundwater flow model are used.

The top row of Figure 5.2 contains the parameters ($S_{\cdot}$ and $P_{\cdot}$) of the stochastic model. The RVs of these parameters define the prior distributions of the model. The second row contains the parameterized distributions (D_i and K_i). Initially, these distributions represent the prior predictive distributions of the layer thickness D_i and the hydraulic conductivity K_i. The other nodes, denoted by a double border, are deterministic nodes. These nodes are the result of an arithmetic operation on their respective parent variables. The variable C, the total vertical hydraulic resistance, is defined as $C = \sum_{i=1}^{n} D_i/K_i$. The observations of node C are denoted by c_{m}.

A problem of the deterministic nodes in such a network is that the joint distribution does not exist [*Cobb and Shenoy*, 2005, 2006; *Cinicioglu and Shenoy*, 2009], and hence the likelihood function, given c_{m}, can strictly spoken not be defined. Only the likelihood functions of the RVs D_i and K_i can be defined, but of these variables no observations are available. In Section 5.1.4 a solution to this problem is described.

5.1.3 Reverse model building

The aim of a Bayesian analysis is to find the most likely distribution, or the parameters of this distribution, of an observable but unknown variable. Bayesian data analysis usually starts with the definition of a full probability model for an observable variable. Such a model consists of a parameterized probability density function (PDF), of a family of density functions, which can describe the uncertainty of an observable quantity. Since only the family of distributions is assumed to be known, and not the parameter values itself, these parameters are described as random variables as well following marginal distributions. These marginal distributions also have an assumed shape. All these distributions together form a joint density function. Integration of this model over the prior marginal distributions yields the so called prior predictive distribution of the observable quantity.

In this chapter, this step of the Bayesian inference is reversed. In our study, we have probability distributions available of the hydraulic conductivity and the layer thicknesses of each litho-class, but no full stochastic model of these variables. These distributions are regarded as the prior predictive distributions of their respective variables. To arrive at a full stochastic model, each prior predictive distribution is decomposed into a stochastic model with PDFs of the marginal distributions of the parameters. This does not necessarily yield any standard distribution for the model or the parameters.

In this section, the method is described as a general method. Therefore, the used symbols do not coincide with the variables of the overall problem described in this chapter. Hereafter, the variable Y represents the prior predictive distribution, and can be read as either D_i or K_i, which are defined in Section 5.1.2.

Data model definition

Let Y be a random variable (RV), with y being the observed data, with prior predictive distribution $f(y)$, and let θ be a vector of parameters of the data model $f_Y(y|\theta)$. The data model also is the likelihood [*Andreon and Weaver*, 2015, p. 22], and also called data distribution or sampling distribution [*Gelman et al.*, 2014, p. 6]. The vector θ is of size n with $n \geq 1$. In a Bayesian context, the parameters θ are uncertain, which uncertainty is described by the probability density function (PDF) $f(\theta)$. The PDF $f(\theta)$ is the prior distribution of θ. As mentioned above, a Bayesian analysis usually starts with the definition of $f(\theta)$ and $f_Y(y|\theta)$. Subsequently, the prior predictive distribution $f(y)$ is obtained by integrating θ out, which writes

$$f(y) = \int_\theta f_Y(y|\theta) f(\theta) \,\mathrm{d}\theta. \tag{5.1}$$

Here, this part is reversed. It is assumed that the prior predictive distribution $f(y)$ is known. To define the data model $f_Y(y|\theta)$, an unknown density function $f(x)$ is defined. Instead of defining $f(x)$ as a function parameterized by θ, a transformation function is defined which describes the relation between the variables x, θ and y. The PDF of X is called the base function of the data model [*Kroese et al.*, 2011,

p. 48]. The transformation function is defined as

$$y = g_x(x, \theta), \tag{5.2}$$

with $g_x(\cdot)$ being a strictly monotone function, with respect to x, within the domain of its parameters θ and x. For the problem at hand, function $g_x(\cdot)$ yields the same result whether it is strictly decreasing or strictly increasing. Hereafter, the transformation function is assumed to be strictly increasing. Equivalently, the inverse function $x = g_y(y, \theta)$ is defined, which must exist. Applying the change of variables [e.g. *Held and Bové*, 2013, p. 321], the data model can be defined as

$$f_Y(y|\theta) = \frac{f(x)}{|g'_x(x, \theta)|} = f_X(g_y(y, \theta))|g'_y(y, \theta)|, \tag{5.3}$$

where $g'_x(x, \theta)$ is the first derivative of $g_x(x, \theta)$ with respect to x, and $g'_y(y, \theta)$ is the first derivative of $g_y(y, \theta)$ with respect to y. The prior predictive distribution of Equation (5.1) may now be written as

$$f(y) = \int_\theta f_X(g_y(y, \theta))|g'_y(y, \theta)| f(\theta)\, \mathrm{d}\theta. \tag{5.4}$$

Since we stated that the $g_x(\cdot)$ is strictly increasing, both derivatives $g'_x(\cdot)$ and $g'_y(\cdot)$ have positive values. So the absolute bars in Equation (5.3) can be omitted in Equation (5.4).

Still, only the prior predictive distribution $f(y)$ is known and not the PDFs of X and θ. In the next section, an iterative method is explained to decompose the PDF of Y into the distributions of the data model, given a transformation function.

Transformation and marginalization

Calculation of marginal distributions of the probability model is an important part of Bayesian modeling. This requires integration over these parameters, but the solutions of these integrals are not always easily achieved. In the former section, the use of a base function $f(x)$ is proposed, combined with transformation function $g_x(\cdot)$. In this study, we have chosen for a location-scale transformation function [*Kroese et al.*, 2011, p. 47] with location parameter (θ_1) and scale parameter (θ_2), which are related to the mean and standard deviation of the distribution, respectively. Since $f(x)$ is not parameterized by θ, the function $f(x)$ has a fixed shape. This is contrary to the usual definition of a location-scale model, where $f(x)$ is parameterized by θ. Therefore, the used model will hereafter be called a location-scale-shape model. So the transformation function writes

$$y = g_x(x, \theta) = \theta_1 + \theta_2 x, \tag{5.5}$$

with its first derivative with respect to x being $g'_x(x, \theta) = \theta_2$. This transformation function describes the deterministic relation between the variables, but the same expression can also be used to describe a combination of the respective random variables (RVs) as

$$Y = \Theta_1 + \Theta_2 X, \tag{5.6}$$

where Θ_1 and Θ_2 are the RVs describing the uncertainty of θ_1 and θ_2, respectively. The RVs Θ_1, Θ_2 and X are defined on their respective finite domains $[\theta_1^{\min}, \theta_1^{\max}]$, $[\theta_2^{\min}, \theta_2^{\max}]$, and $[x^{\min}, x^{\max}]$. The requirement for finite domains is caused by the piecewise linear description of all PDFs. Variable Y is known, having any arbitrary proper probability density function. The variables Θ_1, Θ_2, and X have to be determined by a decomposition of Y into these three variables. Since Y and X are not independent, the calculation of X, given Θ_1 and Θ_2, is not straight forward and is therefore achieved by an iterative Monte Carlo algorithm. Details of the application of this algorithm are found in Section 5.2.1. Equation (5.6) is split into two equations

$$Y = \Theta_1 + U \quad \text{and} \quad U = \Theta_2 X. \tag{5.7}$$

First, initial guesses for the distributions of Θ_1 and U are made, Θ_1^0 and U^0, respectively. Herewith, the calculation $\hat{Y} = \Theta_1^0 + U^0$ is performed. Subsequently, the ratio $f_Y(y)/f_{\hat{Y}}(y)$ is used to modify the PDF of U^0. This is iterated until the PDF of $\hat{Y}$ has, to some rate, converged to the PDF of Y. The result is a new guess for U^0

$$\tilde{U} = U^0 \quad \rightarrow \quad Y \approx \Theta_1^0 + \tilde{U} \quad \rightarrow \quad U^1 = \tilde{U}. \tag{5.8}$$

Second, initial guesses for the distributions of Θ_2 and X are made, Θ_2^0 and X^0, respectively. This yields in an equivalent way updated PDFs of Θ_2 and X as

$$\tilde{\Theta}_2 = \Theta_2^0 \quad \rightarrow \quad U^1 \approx \tilde{\Theta}_2 X^0 \quad \rightarrow \quad \Theta_2^1 = \tilde{\Theta}_2, \tag{5.9}$$

and

$$\tilde{X} = X^0 \quad \rightarrow \quad U^1 \approx \Theta_2^1 \tilde{X} \quad \rightarrow \quad X^1 = \tilde{X}. \tag{5.10}$$

Finally, Θ_1 is estimated again with a fixed value of U and Y as

$$\tilde{\Theta}_1 = \Theta_1^0,\ U^2 = \Theta_2^1 X^1 \quad \rightarrow \quad Y \approx \tilde{\Theta}_1 + U^2 \quad \rightarrow \quad \Theta_1^1 = \tilde{\Theta}_1. \tag{5.11}$$

In this algorithm, it is an advantage to describe all involved PDFs by piecewise linear functions. The modification of the PDFs during the iterative process will most probably never yield a distribution of some standard distribution family. Therefore, no standard (analytical) solutions for the binary operations on the random variables, like summation and multiplication, are available. By using piecewise linear approximations of the PDFs, these calculations can be performed regardless of the shape of the functions.

5.1.4 Inference in the Bayesian Network

As mentioned before, in a Bayesian inference an important aim is to find the posterior distributions of the parameters of a stochastic model given the observations. In this section, the inference of the posterior distributions from a stochastic model with respect to observations at a deterministic node is described.

Joint distribution

In Figure 5.2 a Bayesian network with stochastic and deterministic nodes is depicted, showing the problem at hand in this chapter. Each node represents a random variable (RV). The variables of interest are the layer thickness D_i and the hydraulic conductivity K_i. These variables are described by a parameterized probability density function (PDF) with location parameters $P_{\cdot}$ and scale parameters $S_{\cdot}$. These parameters are uncertain too and are described by their respective PDFs. The network of Figure 5.2 is used to derive the posterior distributions of the parameters $P_{\cdot}$ and $S_{\cdot}$, given an observation of C.

First, the joint distributions are defined. The presented network contains stochastic and deterministic nodes. Unfortunately, a joint PDF including deterministic nodes can not be defined [*Cobb and Shenoy*, 2005, 2006; *Cinicioglu and Shenoy*, 2009]. Therefore, we first define the joint PDF of the stochastic nodes and implement the deterministic relations thereafter. For conciseness and readability, the parameters of the distributions of D_i and K_i are defined as $\theta_{d_i} = \{p_{d_i}, s_{d_i}\}$ and $\theta_{k_i} = \{p_{k_i}, s_{k_i}\}$, respectively. Subsequently, the set of all parameters of C_i is $\theta_i = \{\theta_{d_i}, \theta_{k_i}\}$, and the set of all parameters of the network is $\theta = \{\theta_1, \ldots, \theta_n\}$, with n being the number of litho-classes. The conditional joint distributions of the stochastic nodes, the layer thickness D_i and conductivity K_i, with their respective parameters θ_{d_i} and θ_{k_i}, write

$$f(d_i, \theta_{d_i}) = f(d_i|\theta_{d_i}) f(\theta_{d_i}) \tag{5.12}$$

$$f(k_i, \theta_{k_i}) = f(k_i|\theta_{k_i}) f(\theta_{k_i}). \tag{5.13}$$

The joint PDF of all the stochastic variables in the network writes

$$f(d, k, \theta) = f(d, k|\theta) f(\theta) = \prod_{i=1}^{n} f(d_i|\theta_{d_i}) f(k_i|\theta_{k_i}) f(\theta_{k_i}) f(\theta_{d_i}), \tag{5.14}$$

with $d = \{d_1, \ldots, d_n\}$ and $k = \{k_1, \ldots, k_n\}$. All variables D_i and K_i and their respective location and scale parameters are assumed to be mutually independent.

In this study, we aim to decrease the uncertainty of the variables D_i and K_i by decreasing their parameter uncertainties (θ). This is conducted by adopting the Bayesian approach. According to Bayes' theorem, the posterior distribution of the parameters θ writes

$$f(\theta|d, k) = \frac{f(\theta, d, k)}{\int f(\theta, d, k)\,\mathrm{d}\theta} = \frac{f(d, k|\theta) f(\theta)}{f(d, k)}, \tag{5.15}$$

with d and k observed. Unfortunately, in the problem at hand we do not have observations of d and k. Moreover, if we had observations of each layer thickness and conductivity, it would not have been necessary to define the joint distribution of all layers, but only the distributions for each variable D_i or K_i. Here, only an observation of the total vertical resistance $C = c_{\mathrm{m}}$ is available. But C is a deterministic variable and its distribution is no part of the joint PDF (the numerator of

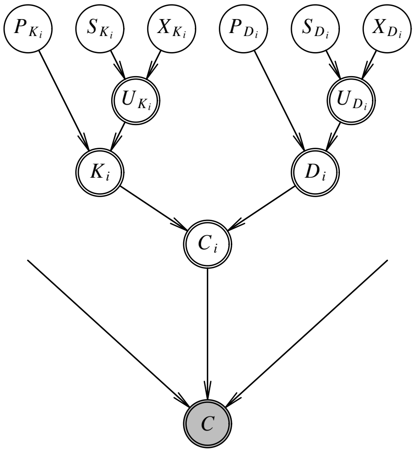

Figure 5.3: Bayesian graph with the models of the nodes K_i and D_i redefined compared to Figure 5.2. Only the marginal nodes are now stochastic, all other nodes are deterministic.

Equation (5.15)). Thereby, making inference for $4n$ parameters, which is the size of θ, may be intractable. Therefore, the Bayesian network has to be reorganized to achieve a workable expression of the equations, and it should be able to implement the deterministic dependencies.

Deterministic variables

As said before, a joint PDF including deterministic nodes can not be defined [*Cobb and Shenoy*, 2005, 2006; *Cinicioglu and Shenoy*, 2009]. Nevertheless, methods are available to handle Bayesian networks with deterministic variables.

In Figure 5.3, the variables K_i and D_i are redefined compared to Figure 5.2. In this figure, the models of K_i and D_i are expanded into a location-scale-shape model according to the method as described in Section 5.1.3. Herein is $X_.$ a variable which describes the shape or family of the corresponding distribution, and are K_i and D_i deterministically defined as $K_i = P_{K_i} + S_{K_i} X_{K_i}$ and $D_i = P_{D_i} + S_{D_i} X_{D_i}$, respectively. The PDF of X is called the base function [*Kroese et al.*, 2011, p. 48]. By making information of K_i and D_i explicit through variables $X_.$, these variables are now converted from stochastic into deterministic, although the joint distribution functions of these variables do not change. So, it can be seen that the distinction between deterministic and stochastic variables in a Bayesian network is not necessarily very strict.

Still, the deterministic nodes are not included in the joint distribution. To circumvent this problem, the deterministic nodes can be described by Dirac delta-functions for continuous distributions [*Cinicioglu and Shenoy*, 2009] or indicator functions for discrete distributions [*Cobb and Shenoy*, 2006]. In this study, the distributions are assumed to be continuous. In this application of the Dirac delta-function ($\delta(\cdot)$), the function value is 1 when its argument is 0, otherwise the function value is 0. When integrating this function over the interval $(-\infty, \infty)$ the result is 1. The integral function of the Dirac delta-function is often denoted as the Heaviside function. With these definitions, the Dirac delta-function acts as a probability mass functions (PMF), with all its mass located at domain value 0.

Before writing the joint distribution of the complete network, including the deterministic relations, the deterministic relations of the network and the conditional distributions with the corresponding Dirac delta-functions have to be defined. These write

$$\begin{aligned}
c = c_\mathrm{m} = \textstyle\sum c_i &\quad\rightarrow\quad & f(c|c_1, ..., c_n) &= \delta(c - \textstyle\sum c_i) \\
c_i = d_i/k_i &\quad\rightarrow\quad & f(c_i|d_i, k_i) &= \delta(c_i - d_i/k_i) \\
d_i = p_{D_i} + u_{D_i} &\quad\rightarrow\quad & f(d_i|p_{D_i}, u_{D_i}) &= \delta(d_i - (p_{D_i} + u_{D_i})) \\
u_{D_i} = s_{D_i} x_{D_i} &\quad\rightarrow\quad & f(u_{D_i}|s_{D_i}, x_{D_i}) &= \delta(u_{D_i} - s_{D_i} x_{D_i}) \\
k_i = p_{K_i} + u_{K_i} &\quad\rightarrow\quad & f(k_i|p_{K_i}, u_{K_i}) &= \delta(k_i - (p_{K_i} + u_{K_i})) \\
u_{K_i} = s_{K_i} x_{K_i} &\quad\rightarrow\quad & f(u_{K_i}|s_{K_i}, x_{K_i}) &= \delta(u_{K_i} - s_{K_i} x_{K_i}).
\end{aligned} \tag{5.16}$$

Herewith, Equation (5.14) can be rewritten including the functions of these new variables. For readability ϕ is defined as the set of marginal distributions, with $\phi = \{\phi_1, \ldots, \phi_n\}$, $\phi_i = \{\phi_{K_i}, \phi_{D_i}\}$, $\phi_{K_i} = \{p_{K_i}, s_{K_i}, x_{K_i}\}$, and $\phi_{D_i} = \{p_{D_i}, s_{D_i}, x_{D_i}\}$. Herewith, the equation of the Bayesian network writes

$$\begin{aligned}
&f(c, c_i, k_i, d_i, u_{K_i}, u_{D_i}, \phi_i; i = 1, \ldots, n) = \\
&\qquad f(c|c_1, ..., c_n) \prod_{i=1}^{n} f(c_i|d_i, k_i) f(k_i|p_{K_i}, u_{K_i}) f(d_i|p_{D_i}, u_{D_i}) \\
&\qquad\qquad f(u_{K_i}|s_{K_i}, x_{K_i}) f(u_{D_i}|s_{D_i}, x_{D_i}) f(\phi_i),
\end{aligned} \tag{5.17}$$

or

$$\begin{aligned}
&f(c, c_i, k_i, d_i, u_{K_i}, u_{D_i}, \phi_i; i = 1, \ldots, n) = \\
&\qquad \delta(c - \textstyle\sum c_i) \displaystyle\prod_{i=1}^{n} \delta(c_i - d_i/k_i) \delta(k_i - (p_{K_i} + u_{K_i})) \delta(d_i - (p_{D_i} + u_{D_i})) \\
&\qquad\qquad \delta(u_{K_i} - s_{K_i} x_{K_i}) \delta(u_{D_i} - s_{D_i} x_{D_i}) f(\phi_i),
\end{aligned} \tag{5.18}$$

with only the functions of the marginal distributions pertaining to stochastic variables. Naturally, this increases the size of the expression, whereas the size should decrease to make inferences feasible. Decreasing the size of the expression can be done when making inference of only a limited number of variables at a time. A standard way of integrating out a deterministic node from a joint distribution of a deterministic node and stochastic nodes [*Khuri*, 2004; *Cinicioglu and Shenoy*, 2009] is

$$\begin{aligned}
f(y) = \int \delta(y - g_x(x)) f(x) \, \mathrm{d}x &= \int \delta(g_y(y) - x) f(x) \, \mathrm{d}x \\
&= \left| \frac{\mathrm{d}g_y(y)}{\mathrm{d}y} \right| f(g_y(y)),
\end{aligned} \tag{5.19}$$

with the deterministic relations $y = g_x(x)$, $x = g_y(y)$ and $z = h_x(x)$ and $x = h_z(z)$, where the functions g and h must exist and g_y is differentiable. In this expression $f(x)$ may represent a joint distribution with x of size ≥ 1. This integration is an application of a change of variables [e.g. *Held and Bové*, 2013, p. 321], and is hereafter applied in the marginalization of the deterministic variables.

Integrating out, or marginalization, of the nodes to which currently no inference is made, is discussed in the next section.

Marginalization and Likelihood

When making inference in a Bayesian network, one may be not interested in an update of all (marginal) distributions but only in a limited number, which may yield a tractable formulation of the problem. On the other hand, when one is interested in an update of all marginal distributions, the complete joint distribution is often too large to make this inference at once. In such cases, the variables which are, temporarily, of no interest are marginalized or integrated out.

In Figure 5.2 the total graph of the problem at hand is shown, and in Figure 5.3 a redefinition of the nodes K_i and D_i is depicted. This graph contains $6n$ (n being the number of litho-classes) marginal distributions, which is too many to update all these distributions at once, given an observation of C. The theory allows to perform the inference for one marginal distribution at the time. If the inference is performed for one variable, say variable φ with $\varphi \in \{P_{K_i}, S_{K_i}, X_{K_i}, P_{D_i}, S_{D_i}, X_{D_i}\}$ and $i \in \{1, \ldots, n\}$, then the calculation of the posterior distributions becomes tractable. Herewith, the joint distribution writes $f(c, \varphi) = f(c|\varphi)f(\varphi) = f(\varphi|c)f(c)$, and the desired posterior distribution of φ writes as Bayes' Theorem

$$f(\varphi|c) = \frac{f(c|\varphi)f(\varphi)}{f(c)}. \tag{5.20}$$

In this expression the likelihood function $f(c|\varphi)$ is unknown and has to be derived by marginalization. Consecutively, the prior distribution $f(\varphi)$ can be updated to the posterior distribution $f(\varphi|c)$, which is the aim of the inference.

Since here all PDFs are described by piecewise linear functions, the marginalization in the direction of the arrows over the deterministic nodes is easily performed, regardless of the shape of the PDFs (see Chapter 2). If we are, for the moment, only interested in the posterior distributions of litho-class i, all distributions of the other litho-classes can be marginalized out by applying arithmetic operations on the RVs of the other litho-classes. This can be written as

$$C_J = \sum_{j \in J} (P_{D_j} + S_{D_j} X_{D_j})/(P_{K_j} + S_{K_j} X_{K_j}), \tag{5.21}$$

with J being the set of all indices excluded index i, thus $J = \{1, \ldots, n\} \setminus \{i\}$. After this operation, C_J is the sum of all vertical resistances of litho-classes j with $j \in J$. The node C_J contains now a marginal distribution and is therefore a stochastic node now. The result of this marginalization is shown in Figure 5.4a. The variable

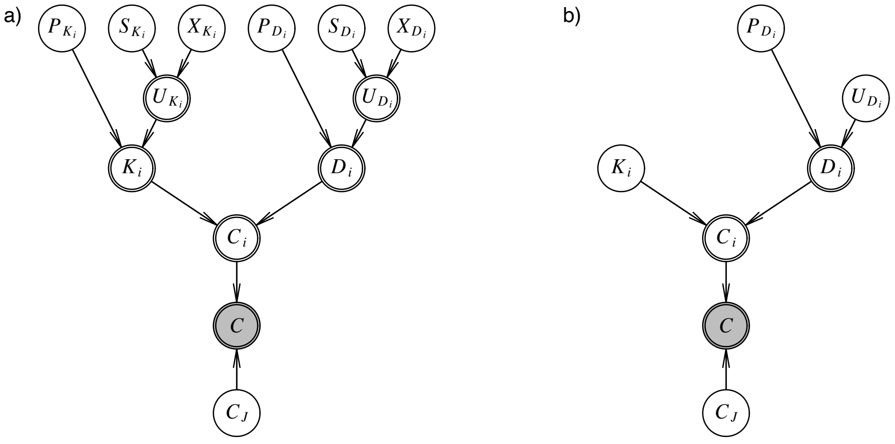

Figure 5.4: *Figure a) shows the same graph as Figure 5.3 but with the variables of all litho-classes, except class i, marginalized out into variable C_J. Figure b) shows further marginalization of the marginal nodes of C_i, except variable of interest P_{D_i}.*

C_i has still six marginal distributions. If we pick one of these variables, e.g. the location parameter P_{D_i}, further simple marginalization (applying arithmetic operations on piecewise linear PDFs) yields Figure 5.4b. After these operations, Eq. 5.18 is reduced to

$$\begin{aligned} f(c, c_J, c_i, k_i, d_i, u_{D_i}, p_{D_i}) =& \delta(c - (c_J + c_i)) f(c_J) \delta(c_i - d_i/k_i) f(k_i) \\ & \delta(d_i - (p_{D_i} + u_{D_i})) f(u_{D_i}) f(p_{D_i}) \\ =& f(c|c_i, c_J) f(c_J) f(c_i|d_i, k_i) f(k_i) \\ & f(d_i|p_{D_i}, u_{D_i}) f(u_{D_i}) f(p_{D_i}), \end{aligned} \tag{5.22}$$

which also may be written as Bayes' theorem in Equation (5.20) as

$$f(p_{D_i}, \phi|c) = \frac{f(c|p_{D_i}, \phi) f(p_{D_i})}{f(c)}, \tag{5.23}$$

with $\phi = \{c_J, c_i, k_i, d_i, u_{D_i}\}$ being the variables to be integrated out. This writes in short

$$\begin{aligned} \int_\phi f(p_{D_i}, \phi|c) \, \mathrm{d}\phi &= f(p_{D_i}|c) \\ &= \frac{f(p_{D_i}) \int_\phi f(c|p_{D_i}, \phi) \, \mathrm{d}\phi}{f(c)} \\ &= \frac{f(p_{D_i}) f(c|p_{D_i})}{f(c)}, \end{aligned} \tag{5.24}$$

with the to be acquired likelihood function

$$\begin{aligned} f(c|p_{D_i}) &= \int_\phi f(c|p_{D_i},\phi)\,\mathrm{d}\phi \\ &= \int_\phi \delta(c-(c_J+c_i))f(c_J)\delta(c_i-d_i/k_i)f(k_i) \\ &\quad\ \delta(d_i-(p_{D_i}+u_{D_i}))f(u_{D_i})\,\mathrm{d}\phi. \end{aligned} \tag{5.25}$$

At this point, arc-reversal [*Shachter*, 1986; *Shachter*, 1988; *Cinicioglu and Shenoy*, 2009; *Kjærulff and Madsen*, 2012, p. 56,116] could be an option for further inference. The arcs (C_i, C), (D_i, C_i) and (D_i, P_{D_i}) can be reversed to find the posterior distribution of P_{D_i}. However, the required calculations to achieve this are more complicated then the forward deterministic calculations with piecewise linear PDFs. Therefore, no arc-reversal is used hereafter but a method is applied in which the order of calculations is not changed.

The integration of Equation (5.25) can be written as a sequence of binary operations on RVs as

$$C = C_J + (U_{D_i} + p_{Di})/K_i, \tag{5.26}$$

where p_{D_i} has a deterministic value and the other variables are RVs. The PDF of C is now equal to the likelihood function $f(c|p_{D_i})$ for the given value of p_{D_i}. The same result is achieved by applying the method of Equation (5.19) to Equation (5.25) which yields

$$\begin{aligned} f(c|p_{D_i}) = \iint_{c_i,c_J} &\delta(c-(c_J+c_i))f(c_J) \iint_{d_i,k_i} \delta(c_i-d_i/k_i)f(k_i) \\ &\int_{u_{D_i}} \delta(d_i-(p_{D_i}+u_{D_i}))f(u_{D_i})\,\mathrm{d}u_{D_i}\,\mathrm{d}k_i\,\mathrm{d}d_i\,\mathrm{d}c_J\,\mathrm{d}c_i. \end{aligned} \tag{5.27}$$

Subsequent integration yields integrals which are equivalent to the binary arithmetic operations on the RVs. The integral of $D_i = U_{D_i} + p_{D_i}$ writes

$$\begin{aligned} f(d_i|p_{D_i}) &= \int_{u_{D_i}} \delta(d_i-(p_{D_i}+u_{D_i}))f(u_{D_i})\,\mathrm{d}u_{D_i} \\ &= f_{U_{D_i}}(d_i-p_{D_i}). \end{aligned} \tag{5.28}$$

With this result substituted in Equation (5.27), the integral of $C_i = D_i/K_i$ writes

$$\begin{aligned} f(c_i|p_{D_i}) &= \iint_{d_i,k_i} \delta(c_i-d_i/k_i)f(k_i)f(d_i|p_{D_i})\,\mathrm{d}k_i\,\mathrm{d}d_i \\ &= \int_{d_i} \left|\frac{d_i}{c_i^2}\right| f_{K_i}(d_i/c_i)f(d_i|p_{D_i})\,\mathrm{d}d_i \\ &= \int_{d_i} f(c_i|d_i)f(d_i|p_{D_i})\,\mathrm{d}d_i. \end{aligned} \tag{5.29}$$

Subsequent substitution in Equation (5.27) yields the integral of $C = C_J + C_i$ which integrates as

$$\begin{aligned} f(c|p_{D_i}) &= \int_{c_i} f_{C_J}(c - c_i) f(c_i|p_{D_i})\,\mathrm{d}c_i \\ &= \int_{c_i} f(c|c_i) f(c_i|p_{D_i})\,\mathrm{d}c_i. \end{aligned} \tag{5.30}$$

If the function is evaluated for multiple values of p_{D_i}, the likelihood function for P_{D_i} given observations of C can be approximated as a piecewise linear function. The choice of an appropriate discretization of p_{D_i} is described in the next section.

5.1.5 Likelihood function discretization optimization

If parameterized likelihood functions are used then the shape of the function is analytically defined, and for every value in the domain of the function the likelihood is known. In the current application, the likelihood functions are numerically described by piecewise linear functions. A challenging task herein is finding an adequate description of the function, especially when the majority of the probability resides in a small area of its domain [*Kozlov and Koller*, 1997; *Neil et al.*, 2007; *Marquez et al.*, 2010] or when a good approximation of the tails of the distribution is required [*Zhu and Collette*, 2015]. This problem is equivalent to finding an optimal discretization of the results of the calculations with PL-PDFs, as described in Section 2.1.2. There, the calculations are started with three discretization points, and at each point the probability density and the cumulative probability are calculated. Then, between two adjacent discretization points (bin) with the largest discrepancy between the calculated cumulative probability and the probability derived from the linearized probability densities, a new discretization point is added. This is repeated until at each bin the discrepancy is acceptable and an adequate number of discretization points has been calculated.

In case of the likelihood functions this strategy is not applicable, since only one value of a marginal distribution is used, instead of the whole marginal distribution, for the calculation of the likelihood of that specific conditioning marginal value. Therefore, the cumulative probability can not be calculated together with the likelihood density. Nevertheless, it is possible to calculate the likelihood density and its first derivative with respect to the conditioning marginal value. If the discretization of the piecewise linear likelihood function is adequate then the integral of the derivatives at the same discretization should show a good approximation of the likelihood function. This can be calculated for each bin of the discretized function. At bins with a large discrepancy between the two functions the discretization has to be refined.

In fact, a likelihood function is not a PDF since the function not necessarily integrates to one. Therefore, instead of probability density and cumulative probability these values will be denoted by likelihood density and cumulative likelihood, respectively.

Derivative of the likelihood function

As stated above, for the optimization of the discretization of the likelihood function, the likelihood density as well as its derivative with respect to its conditioning marginal value are needed. The calculation of the likelihood function is shown above, the calculation of its derivative is described here.

If the likelihood function $f(c|p_{D_i})$ is used, as defined in Equation (5.27), then its derivative with respect to p_{D_i} is defined as

$$f'(c|p_{D_i}) = \frac{\mathrm{d}f(c|p_{D_i})}{\mathrm{d}p_{D_i}} = \frac{f(c|p_{D_i} + \mathrm{d}p_{D_i}) - f(c|p_{D_i})}{\mathrm{d}p_{D_i}}, \tag{5.31}$$

for $\lim \mathrm{d}p_{D_i} \to 0$. Under the same conditions, this derivative may also be written as

$$f'(c|p_{D_i}) = \frac{[f(c|p_{D_i}) + \mathrm{d}p_{D_i} f'(c|p_{D_i})] - f(c|p_{D_i})}{\mathrm{d}p_{D_i}}. \tag{5.32}$$

This can be applied to Equation (5.27) with substitution of Equation (5.28). If we define for readability

$$f(c, c_J, c_i, k_i|d_i) = \delta(c - (c_J + c_i)) f(c_J) \delta(c_i - d_i/k_i) f(k_i), \tag{5.33}$$

then the likelihood function writes

$$f(c|p_{D_i}) = \underset{c_i, c_J, d_i, k_i}{\int \cdots \int} f(c, c_J, c_i, k_i|d_i) f(d_i|p_{D_i}) \,\mathrm{d}k_i \,\mathrm{d}d_i \,\mathrm{d}c_J \,\mathrm{d}c_i, \tag{5.34}$$

with only the function $f(d_i|p_{D_i}) = f_{U_{D_i}}(d_i - p_{D_i})$ dependent on p_{D_i}. Herewith, the first term in the numerator of the right hand side of Equation (5.31), and applying its expansion of Equation (5.32), writes

$$\begin{aligned} &f(c|p_{D_i} + \mathrm{d}p_{D_i}) = \\ &\quad \underset{c_i, c_J, d_i, k_i}{\int \cdots \int} f(c, c_J, c_i, k_i|d_i) \left[f(d_i|p_{D_i}) + \mathrm{d}p_{D_i} f'(d_i|p_{D_i}) \right] \mathrm{d}k_i \,\mathrm{d}d_i \,\mathrm{d}c_J \,\mathrm{d}c_i. \end{aligned} \tag{5.35}$$

Applying this to Equation (5.32) yields, in accordance with Leibniz's rule,

$$f'(c|p_{D_i}) = \underset{c_i, c_J, d_i, k_i}{\int \cdots \int} f(c, c_J, c_i, k_i|d_i) f'(d_i|p_{D_i}) \,\mathrm{d}k_i \,\mathrm{d}d_i \,\mathrm{d}c_J \,\mathrm{d}c_i, \tag{5.36}$$

where

$$f'(d_i|p_{D_i}) = \frac{\mathrm{d}f_{U_{D_i}}(d_i - p_{D_i})}{\mathrm{d}u_{D_i}} \frac{\mathrm{d}u_{D_i}}{\mathrm{d}d_i} \frac{\mathrm{d}d_i}{\mathrm{d}p_{D_i}} = -f'_{U_{D_i}}(d_i - p_{D_i}), \tag{5.37}$$

with $f'_{U_{D_i}}$ being the derivative of $f_{U_{D_i}}$ with respect to u_{D_i}. Obviously, this function is not a PDF anymore, but the integral did not change. Therefore, the same algorithm as applied on the binary operations with PDFs can be used to calculate the results,

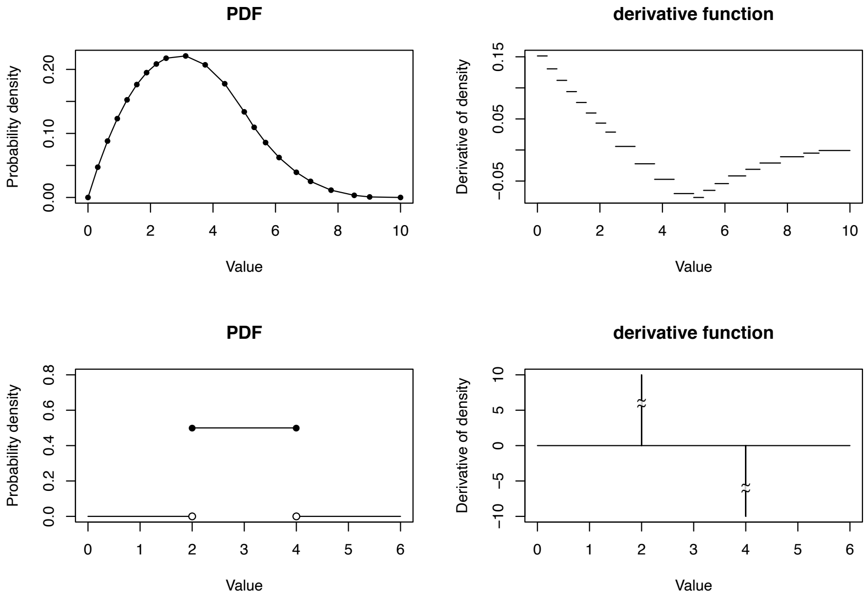

Figure 5.5: Example of a piecewise linear PDF *(left) and its derivative function (right). In the top row the* PDF *is a continuous function and in the bottom row discontinuous. The latter has singular points in its derivative.*

with the addition that negative function values have to be supported. If the variable D_i', from the above example, is attributed with the function $f'(d_i|p_{D_i})$ then the next calculations are performed to find the derivative of the likelihood function $f'(c|p_{D_i})$

$$C_i' = D_i'/K_i \tag{5.38}$$

$$C' = C_J + C_i', \tag{5.39}$$

where the superscript quotes denote that the functions of the variables are derivatives instead of PDFs. These last two equations are equal to the integral of Equation (5.36). This shows how to apply the same algorithms as used for the calculation of the deterministic variables in the Bayesian network like in Equation (5.26).

Singular points in derivatives

The derivative function of a piecewise linear PDF has a constant value within each bin, which is the slope of the linear function within this bin. In Figure 5.5 an example is shown. In the top row of this figure, a PDF without discontinuities is depicted. This yields a derivative function with finite function values (top right pane). In the bottom row, a PDF of a uniform distribution is shown. At the begin and the end of the area of the PDF where the density is greater than zero, the probability density is discontinuous. These discontinuities yield singular points in the derivatives of which the values tend to $\pm\infty$. Nevertheless, these singularities are of importance and must be part of the integral of Equation (5.36). The singularities do not occur in all situations but are still not a rare phenomena. If in this example of the uniform distribution the singular points of the derivative function are ignored, it is obvious that the integration of this derivative function never yields the uniform distribution back again. In the example of Equation (5.36) in the former section the differentiation is not with respect to the same variable as the integration, but for the same reason the singular points can not be ignored. Integration of the singular points is, at least numerically, an infeasible problem. Hereafter is shown how this problem is circumvented.

The derivative of $f(d_i|p_{D_i})$ with respect to p_{D_i} at a discontinuity d_s can, analogous to Equation (5.31), be written as

$$f'_{disc}(d_s|p_{D_i}) = \frac{\lim_{d_i \downarrow d_s} f(d_i|p_{D_i}) - \lim_{d_i \uparrow d_s} f(d_i|p_{D_i})}{\mathrm{d}p_{D_i}} = \frac{\Delta f(d_s|p_{D_i})}{\mathrm{d}p_{D_i}}. \tag{5.40}$$

If $\Delta f()$ is written as a function of $f_{U_{D_i}}$, like in Equation (5.37), this yields

$$\Delta f(d_i|p_{D_i}) = \lim_{u_{D_i} \uparrow d_i - p_{D_i}} f_{U_{D_i}}(u_{D_i}) - \lim_{u_{D_i} \downarrow d_i - p_{D_i}} f_{U_{D_i}}(u_{D_i}), \tag{5.41}$$

in which the upper and lower limits of $f_{U_{D_i}}$ are in reversed order. This yields the negation sign as in Equation (5.37).

Assume that the function $f(d_i|p_{p_{D_i}})$ has m discontinuities $d_{s,l}$ with $l = 1, \ldots, m$, and define the set with all singular points as $\mathcal{S} = \{d_{s,1}, \ldots, d_{s,m}\}$. Herewith, the derivative function writes

$$f'(d_i|p_{D_i}) = \begin{cases} f'_{disc}(d_i|p_{D_i}), & \text{for } d_i \in \mathcal{S}, \\ f'_{cont}(d_i|p_{D_i}), & \text{otherwise,} \end{cases} \tag{5.42}$$

where f'_{cont} is defined for the continuous derivatives, and f'_{disc} at the discontinuities of the PDF. Applied to the integral of Equation (5.29) yields

$$\begin{aligned} f'(c_i|p_{D_i}) &= \int_{d_i} f(c_i|d_i) f'(d_i|p_{D_i})\,\mathrm{d}d_i \\ &= \int_{d_i} f(c_i|d_i)\,[f'_{cont}(d_i|p_{D_i}) + f'_{disc}(d_i|p_{D_i})]\,\mathrm{d}d_i \\ &= \int_{d_i} f(c_i|d_i) f'_{cont}(d_i|p_{D_i})\,\mathrm{d}d_i + \sum_{d_i \in \mathcal{S}} f(c_i|d_i) f'_{disc}(d_i|p_{D_i})\,\mathrm{d}d_i. \end{aligned} \tag{5.43}$$

By substituting f'_{disc} by Equation (5.40), the equation rewrites

$$f'(c_i|p_{D_i}) = \int_{d_i} f(c_i|d_i) f'_{cont}(d_i|p_{D_i})\,\mathrm{d}d_i + \sum_{d_i \in \mathcal{S}} f(c_i|d_i) \Delta f(d_i|p_{D_i}) \frac{\mathrm{d}d_i}{\mathrm{d}p_{D_i}}. \tag{5.44}$$

With $d_i = p_{D_i} + u_{D_i}$ (Equation (5.16)) the derivative $\mathrm{d}d_i/\mathrm{d}p_{D_i} = 1$, so the equation reduces to

$$f'(c_i|p_{D_i}) = \int_{d_i} f(c_i|d_i) f'_{cont}(d_i|p_{D_i})\,\mathrm{d}d_i + \sum_{d_i \in \mathcal{S}} f(c_i|d_i) \Delta f(d_i|p_{D_i}). \tag{5.45}$$

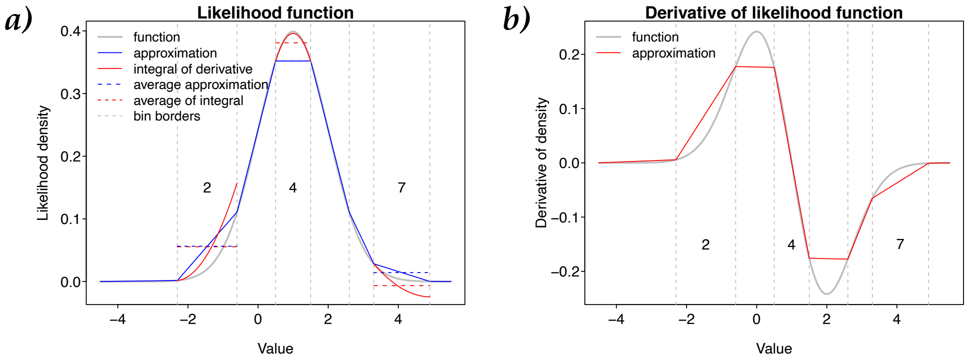

Figure 5.6: Example of a discretized likelihood function (a) and its derivative (b). The numbers are the bin-numbers as used in the text. The outermost edges of bin 1 and 8 are beyond the margins of the figure and therefore not shown.

Herewith, the singular points end up in a (finite valued) summation and are added to the result of the integral of the continuous part of the function.

Bin selection algorithm

Discretization of the likelihood function starts with three discretization points (two bins). The two outermost points are chosen in such a way that beyond these points no significant density is expected. Since the piecewise linear functions have a finite domain, these outermost values are almost always known. Subsequently, iteratively one bin is chosen to add a discretization point. As in the former section described, at each discretization point the likelihood density and its derivative is calculated. Both functions are approximated to be piecewise linear functions. The differences between the integral of the derivative function and the likelihood density function at each bin are calculated. The bin with the largest discrepancy has to be split up into two bins. However, the difference between the likelihood density and the integral of its derivative is not always a good measure for the goodness of the piecewise linear approximation. Therefore, three complementary measures are defined which are described hereafter.

In Figure 5.6 an example with eight bins is given, showing three different measures of discrepancy. Herein, the functions (gray) and their respective piecewise linear approximations are shown. Within each bin, the integral of the derivative function is calculated, which is shown as a solid red line in Figure 5.6a. Also the average density of the likelihood function, and of the integral of the derivative function are shown as a blue and a red dashed line, respectively. The area between two dashed lines is the difference in probability calculated by the two approximations. Now we have three measures available to judge the accuracy of the approximations. First, the difference between the densities (blue and red solid line) at the right side of each bin. In bin 4, these points coincide but in bin 2 and 7 the values differ with the largest discrepancy in bin 2. Second, the difference in average density is used as a measure. In bin 2 this difference is (almost) 0, but is in bin 4 and

7 significant with the largest value in bin 4. Third, the difference in probability is used. In bin 2 this is 0 again, but in bin 4 and 7 this is significant with the largest value in bin 7. Not one of these measures is in every situation distinctive, so these three measures are alternated when selecting a next bin to be split up.

5.1.6 Likelihood with Uncertain observations

In this section, an expression for the likelihood function with uncertain observations is derived. In *Denœux* [2013] and *Denœux* [2014] such an expression is given, but hereafter we arrive at a different formulation which is more appropriate in our case.

Let $f(x)$ be the unknown probability density function (PDF) of X, and let $f(x|\theta)$ be the parameterized PDF of $f(x)$. In a Bayesian context, the PDFs of θ are the prior distributions of θ. Let $\mathcal{L}(\theta|X)$ be the likelihood function of θ, where $X = \{x_1, \ldots, x_n\}$ are n random observations of $f_X(\cdot)$. The likelihood function can now be written as [*Held and Bové*, 2013, p. 18]

$$\mathcal{L}(\theta|x_1, \ldots, x_n) = f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta). \tag{5.46}$$

Assume that n is very large and that x is discrete or that the values of x can be assigned to m classes. The values of these classes are denoted as $\bar{x}_k$, and the number of observations in class k is n_k, hence $n = \sum_{k=1}^{m} n_k$. According to the definition of likelihood functions, Equation (5.46) can be written as

$$\mathcal{L}(\theta|x_1, \ldots, x_n) = \prod_{k=1}^{m} \left(\prod_{j=1}^{n_k} f(\bar{x}_k|\theta) \right) = \prod_{k=1}^{m} \left[f(\bar{x}_k|\theta) \right]^{n_k}. \tag{5.47}$$

With n very large, a probability mass function (PMF) of the observations can be written as

$$p(\bar{x}_k) = \frac{n_k}{n}. \tag{5.48}$$

Equation (5.47) can be written, using the expression of Equation (5.48) $n_k = np(\bar{x}_k)$, as

$$\mathcal{L}(\theta|X) = \prod_{k=1}^{m} \left[f(\bar{x}_k|\theta) \right]^{np(\bar{x}_k)}. \tag{5.49}$$

From Equation (5.49) it can be seen that it does not make any difference if, for every value $\bar{x}$, there are $np(\bar{x})$ observations, or when there are n observations with the complete PMF of $\bar{x}$. So, if only one uncertain observation of $\bar{x}$ is available, described by a PMF, the same equation can be used with $n = 1$. When multiple uncertain observations are available, each described by its own PMF $p_i(\cdot)$, Equation (5.49) can be written as

$$\begin{aligned}\mathcal{L}(\theta|X_{1,\ldots,\nu}) &= \prod_{i=1}^{\nu}\prod_{k=1}^{m}[f(\bar{x}_k|\theta)]^{p_i(\bar{x}_k)} \\ &= \prod_{k=1}^{m}[f(\bar{x}_k|\theta)]^{\sum_{i=1}^{\nu}p_i(\bar{x}_k)}\,,\end{aligned} \tag{5.50}$$

where ν is the number of uncertain observations. If the PMF p_i is smooth, and the class width Δ_x tends to 0, the class probability can be approximated by a PDF f_i as $p_i(\bar{x}_k) = f_i(\bar{x}_k)\Delta_x$. Herewith, Equation (5.50) writes

$$\mathcal{L}(\theta|X_{1,\ldots,\nu}) = \prod_{k=1}^{m}[f(\bar{x}_k|\theta)]^{\sum_{i=1}^{\nu}f_i(\bar{x}_k)\Delta_x}\,. \tag{5.51}$$

Now, we can define the average mixture function of f_i as

$$\bar{f}(x) = \frac{1}{\nu}\sum_{i=1}^{\nu}f_i(x), \tag{5.52}$$

which contains all uncertain information of all observations. Rewriting Equation (5.51) using Equation (5.52) yields

$$\mathcal{L}(\theta|X_{1,\ldots,\nu}) = \prod_{k=1}^{m}[f(\bar{x}_k|\theta)]^{\nu\bar{f}(\bar{x}_k)\Delta_x}\,. \tag{5.53}$$

The log-likelihood can now be written as

$$\ell(\theta|X_{1,\ldots,\nu}) = \nu\sum_{k=1}^{m}\bar{f}(\bar{x}_k)\Delta_x\ln\left(f(\bar{x}_k|\theta)\right). \tag{5.54}$$

With $m \to \infty$ then $\Delta_x \to \mathrm{d}x$, the log-likelihood with uncertain observations in the continuous case writes

$$\ell(\theta|X_{1,\ldots,\nu}) = \nu\int_X \bar{f}(x)\ln\left(f(x|\theta)\right)\mathrm{d}x. \tag{5.55}$$

In Appendix D, the derivation of the likelihood marginalization with uncertain observations is given.

5.2 Results

The methods as described in the former sections are applied to a real-world case of the REGIS hydrogeological model in combination with the calibrated groundwater flow model AZURE [*de Lange and Borren*, 2014]. This study area is described in more detail in Chapter 4. In Figure 5.7 the study area is shown with the total vertical hydraulic resistance of aquitard 4 depicted at the current area of interest. For each litho-class, only one prior probability density function (PDF) is available, without a

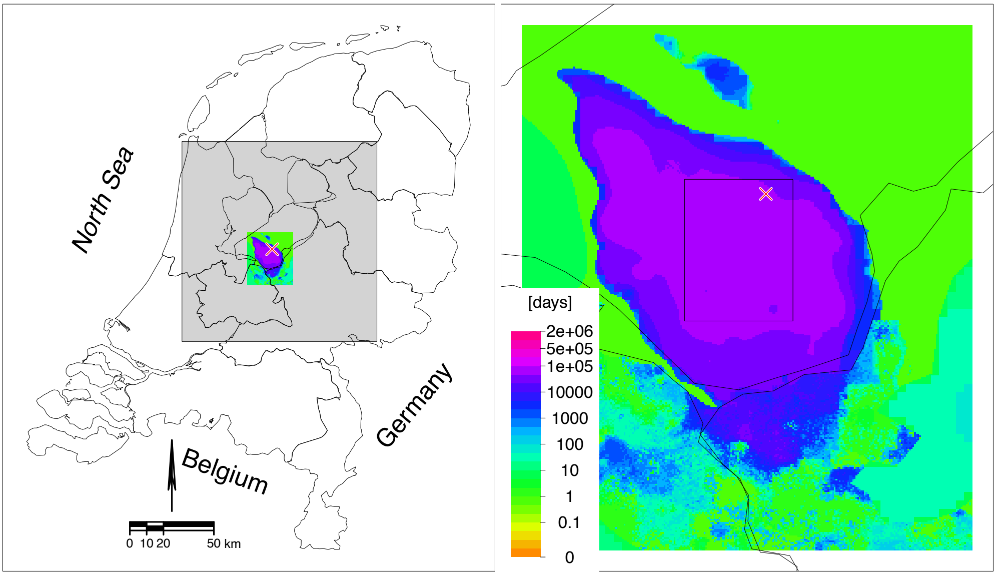

Figure 5.7: Study area in with the total vertical hydraulic resistance of aquitard 4 depicted at the area of interest (left pane). The cross denotes the location of the example in the next section. The gray area denotes the extent of the groundwater flow model from which the calibrated data is used. The right pane shows the enlarged area of interest. The rectangle area delineates the are for which the parameter updates are presented.

lateral differentiation. In Chapter 3 an interpolation of the litho-layer thicknesses is conducted which yields PDFs of the litho-layer thickness at each grid cell of the study area. Those results are used here as the prior distributions of the layer thicknesses.

5.2.1 Stochastic model

In Section 5.1.3 the stochastic models of the variables layer thickness D_i and the vertical hydraulic conductivity K_i are defined generically as $Y = \Theta_1 + \Theta_2 X$ and $U = \Theta_2 X$. Herein is Y the random variable K_i or D_i. In Figure 5.8, an example is given of the results of the evaluation of the stochastic model of one grid cell of the aquitard. The location of this grid cell is denoted in Figure 5.7. To make the figure readable, for the most skew distributions only the most important part is shown, neglecting the tail. The prior distributions (black lines) of the marginal variables Θ_1, Θ_2 and X show the results of the decompositions of the layer thickness (D) and the vertical hydraulic conductivity (K).

The decomposition of the parameters D_i and K_i is performed by the Monte Carlo algorithm as described in Section 5.1.3. This Monte Carlo algorithm starts with arbitrary initial distributions for the marginal variables. The initial distributions of the location parameter Θ_1 and the scale parameter Θ_2 were formed by the summation of three independent uniform distributions. The width of the domain of Θ_1 is chosen as the width around the top of the modeled parameter D_i or K_i

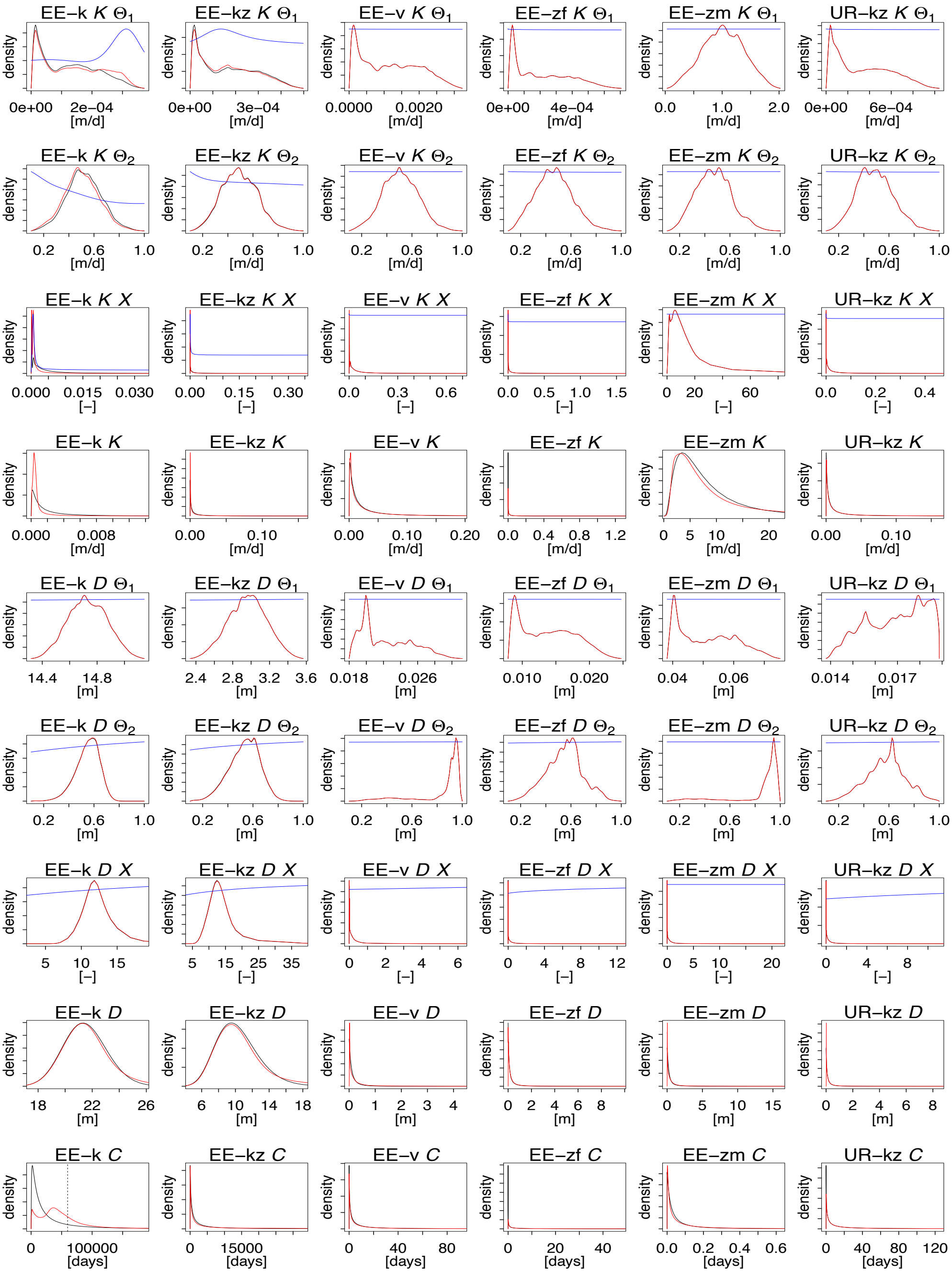

Figure 5.8: *Prior PDFs (black), posterior PDFs (red) and likelihood (blue) functions of a Bayesian update of one gridcell. Each column depicts one litho-class of the marginal variables (Θ_1, Θ_2, X), the layer thickness (D), and the vertical hydraulic conductivity (K) and resistance (C). The dashed black line denotes the value of the observation of the calibrated aquitard resistance. The y-axis units are densities of the PDFs, the likelihood curve is scaled to fit to this axis.*

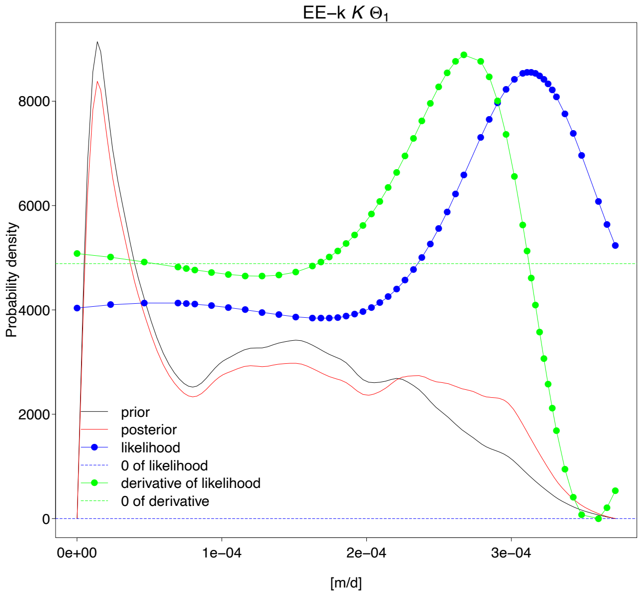

Figure 5.9: *Example of an update of the location parameter Θ_1 of the conductivity K_i of litho-class EE-k. The y-axis is defined for the prior (black) and posterior (red) distributions, the graphs of the likelihood (blue) and its derivative (green) are scaled to fit in the graph. The 0-values of these last graphs are denoted by a horizontal dashed line.*

which describes 0.2 probability of that distribution. The domain of the scale parameter Θ_2 is always set to $[0.1, 1]$, with which the Bayesian update algorithm has the freedom of a factor 10 to scale the distribution of X. The initial distributions of X is chosen as U/Θ_2 (assumed independent) but with its domain adjusted in such a way that the domain of $\Theta_2 X$ equals the domain of U.

The prior distributions of the location parameters Θ_1 for most litho-classes and for variables D_i and K_i differ from their initial distributions, these priors are mostly skewed distributions. The prior distributions of scale parameters Θ_2 of variable K_i are still quite similar to their initial distributions. For all priors holds that the distributions are not very smooth, which is caused by the implementation of the Monte Carlo method. Nevertheless, the distributions of $\Theta_1 + \Theta_2 X$ are smooth and very similar to their original distributions of D_i or K_i.

5.2.2 Posterior distributions

The likelihood functions of all marginal distributions of the Bayesian network are calculated. In Figure 5.9 an example is depicted of the location parameter Θ_1 of the conductivity K_i of litho-class EE-k. The likelihood function (blue line) and its derivative (green line) are determined by the procedure as described in Section 5.1.5. The dots show the iteratively added discretization points of the likelihood function. As desired, the more curved parts of the graph gained more discretization points than the less curved parts. The posterior distribution (red line) is derived by multiplication of the prior by the likelihood and subsequent normalization. In Figure 5.8 it can be seen that for most marginal distributions the likelihood functions (blue) is an almost straight horizontal line, which causes a negligible update of the prior distribution. So most posterior distributions (red) coincide with their respective priors (black). Only the litho-class EE-k has a significant update of its priors. For litho-class EE-kz, marginal distribution X of conductivity K, it can be seen that the iterative algorithm is able to find the likelihood function in very skewed distributions. The domain of X ranges from about 0 to 1.3E4, while the peak of the likelihood function resides between about 0 and 5.E-2. Still, the iterative algorithm found the peak of the likelihood function quite well.

At the bottom row of Figure 5.8 the PDFs of the vertical hydraulic resistance C_i is shown for each layer. Only the posterior distribution of litho-class EE-k diverges noticeable from its prior. The litho-class EE-k has the majority of all the aquitard resistance; all other litho-classes are of minor importance in this example grid cell. Therefore, it is no wonder that the update mainly is assigned to litho-class EE-k. The shape of the posterior distribution of litho-class EE-k is remarkable. Even though the posterior (predictive) distributions of the layer thickness D_i and conductivity K_i are both unimodal, their quotient C_i yields a bimodal distribution. This effect is caused by the shape of the posterior distribution of K_i, which in turn gains this effect by the posterior distribution of X_{K_i}.

5.2.3 Updates of litho-class parameters

For each grid cell of the processed aquitard in the study area, the Bayesian network is evaluated. This yields for each random variable in the network a posterior distribution. The prior distribution of the layer thickness D_i of each litho-class is acquired by kriging interpolation, which yields for each grid cell a different PDF. As an indication of the results of the Bayesian update the mean value and the standard deviation of the prior and posterior distributions are shown in maps of the study area. Also the difference between the posterior and prior value (Δ), and relative adjustment, i.e. the quotient of the posterior and the prior value (ratio), are displayed. For the hydraulic conductivity K_i only one prior distribution for each litho-class is used. Therefore, the prior distribution of K_i has no spatial variation and the maps of these parameters are omitted in the next pictures. For reference, the prior values are displayed in the upper-right corner of each respective posterior variable map.

In Figure 5.10, the update of the aquitard resistance is depicted. The top row shows the mean values of the distributions of the vertical hydraulic resistance μ_C of the aquitard, and the bottom row the accompanying standard deviations. The top-left map shows the calibrated values of the groundwater flow model which are used as observations (obs) in the Bayesian update. The difference between the posterior mean and prior mean ($\Delta\mu_C$) shows that the vertical resistance mostly decreases (blue) or is only slightly changed (yellow). In a small part of the area the resistance increases (red). This is in agreement with the values of the observations compared to the prior means. The maps with ratios (ratio μ_C and ratio σ_C) show

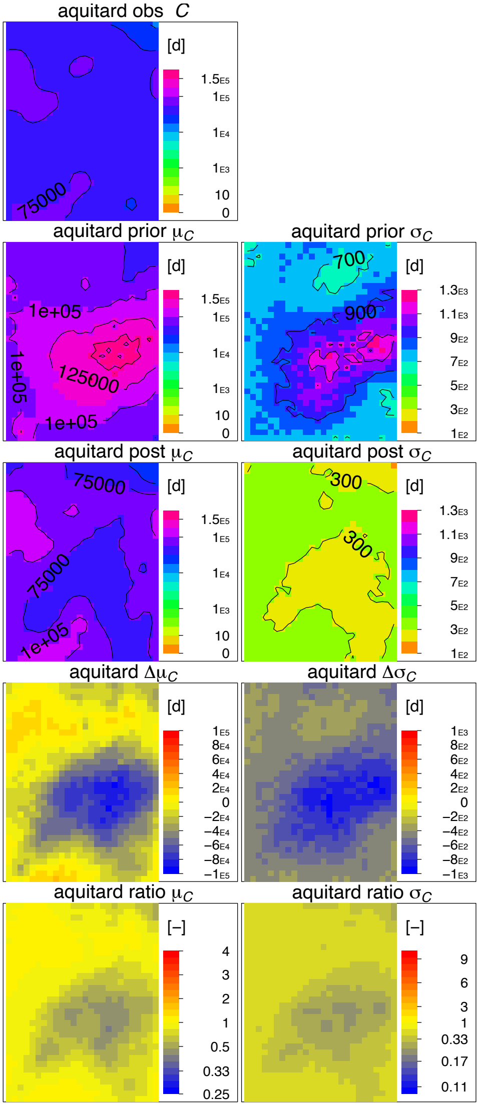

Figure 5.10: *Aquitard vertical hydraulic resistance (days). The top row maps shows the observations (obs), prior and posterior mean (μ_C), the difference between the prior and the posterior mean ($\Delta\mu_C$), and the quotient of the prior and posterior means (ratio μ_C). The bottom row shows the same maps of the standard deviations, except for the observations, of the probability distributions.*

the quotient of the posterior and the prior values. Herein, the posterior vertical resistance is decreased by a factor up to about 0.4, and increased up to a factor of about 1.4. The standard deviations of the posterior distributions are all lower than the standard deviations of the prior distributions. The map with ratios show that the standard deviations are decreased by a factor of about 0.6 to about 0.2.

The legends in the maps of the prior and posterior values are chosen to be equal, for easy comparison of the pictures. The legend of the observations is equal to the legend of the μ_C maps. The legends of the differences maps ($\Delta\mu$ and $\Delta\sigma$) are symmetrical around 0. The legends of the ratios maps (ratio $\Delta\mu$ and ratio $\Delta\sigma$) are symmetrical around 1, which means that the interval values below 1 are the reciprocal of the values above 1.

The aquitard is build-up of deposits of multiple litho-classes. In Figures 5.11 and 5.12 the means and the standard deviations, respectively, of the distributions of the vertical hydraulic resistance C_i are shown for each litho-class separately. Compared to the hydraulic resistance of the classes EE-k (clay) and EE-kz (sandy clay), the other classes do hardly have any influence on the total hydraulic resistance of the aquitard. The resistance of litho-class EE-k is decreased as well as increased significantly, where the resistance of litho-class EE-kz is only decreased or unchanged. The magnitude of the adjustment of litho-class EE-kz is about twice the magnitude of the adjustment of EE-k. In the map of the prior mean resistance of EE-kz (μ_{C_i}, Figure 5.11) an area with resistance is seen. This is clearly caused by two borehole interpretations. In the prior map of EE-k the opposite is shown at the same borehole locations. In the posterior maps these variations almost vanished, and the borehole locations are not clearly expressed anymore. Besides these classes, only litho-class EE-zf (fine sand) has a significant adjustment (decrease) in the resistance up to about 5000 days. This high adjustment is located at a spot with a relative high prior resistance. The standard deviations (Figure 5.12) are decreased for all litho-classes, except for some grid-cells of the litho-class EE-zm (medium sand). This last litho-class is of minor importance since its prior and posterior mean resistances are less than 1 day. For the litho-classes EE-k and EE-kz the factor of maximum adjustment of the standard deviation are about 0.4 and 0.2, respectively. Litho-class EE-zf has even a factor of about 0.12.

Each distribution of the litho-class vertical resistance is a result of the quotient of the random variables of the thickness and the conductivity of the litho-classes. In Figures 5.13 and 5.14 the means and the standard deviations, respectively, of the distributions of the layer thicknesses D_i are shown for each litho-class separately. The third row of Figure 5.13 ($\Delta\mu_{D_i}$) shows that almost all mean layer thicknesses are increased after update, especially in the grid cells of litho-class EE-kz where the prior mean layer thickness is relatively thin. The standard deviations of the two main litho-classes, EE-k and EE-kz, are increased as well up to a factor of about 1.4.

In Figures 5.15 and 5.16 the means and the standard deviations, respectively, of the distributions of the vertical hydraulic conductivity K_i are shown for each litho-

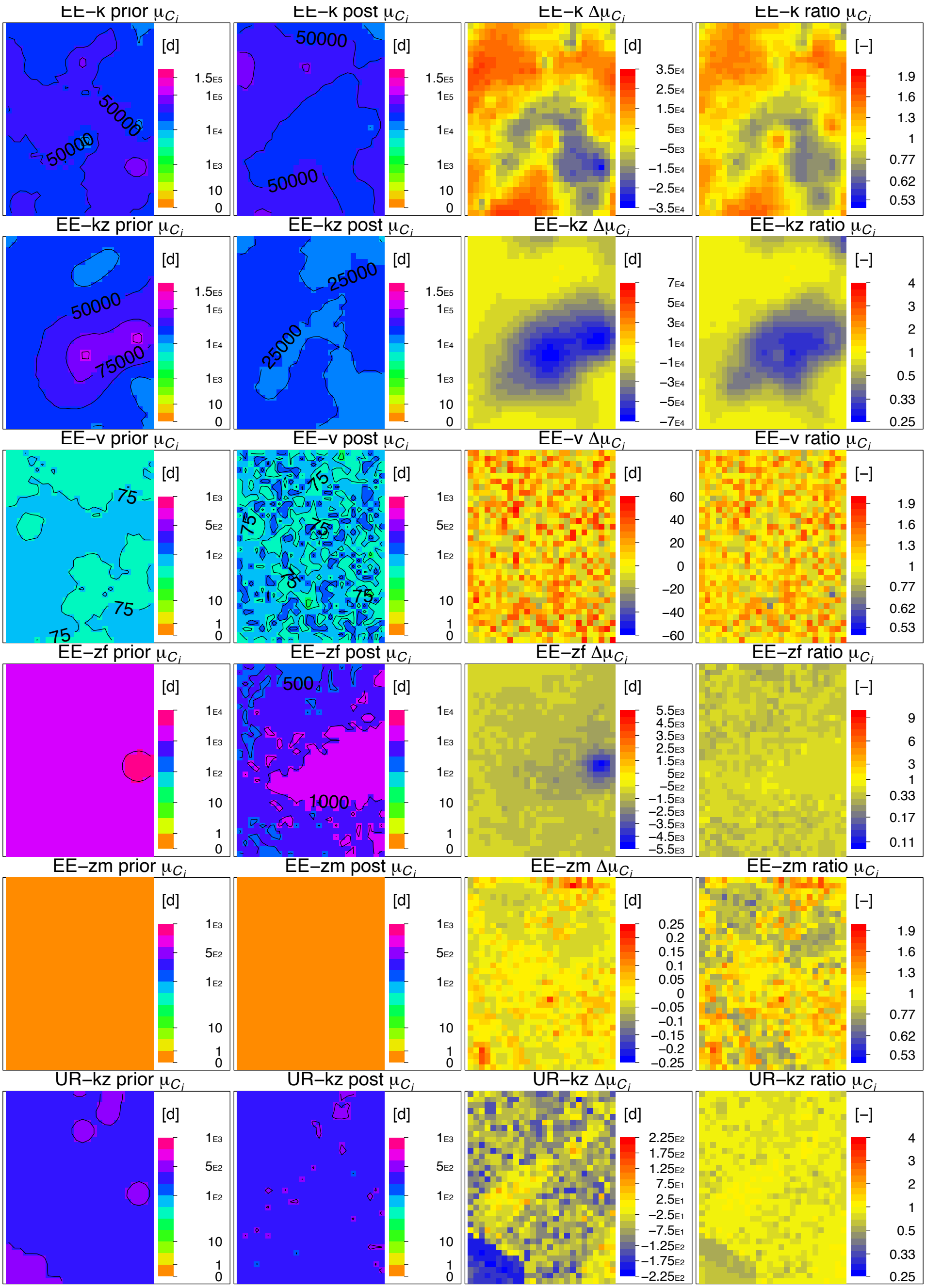

Figure 5.11: *The mean vertical hydraulic resistance C_i for each litho-class is depicted per column. From top to bottom row, the prior and posterior standard deviation, and their difference and ratio are shown.*

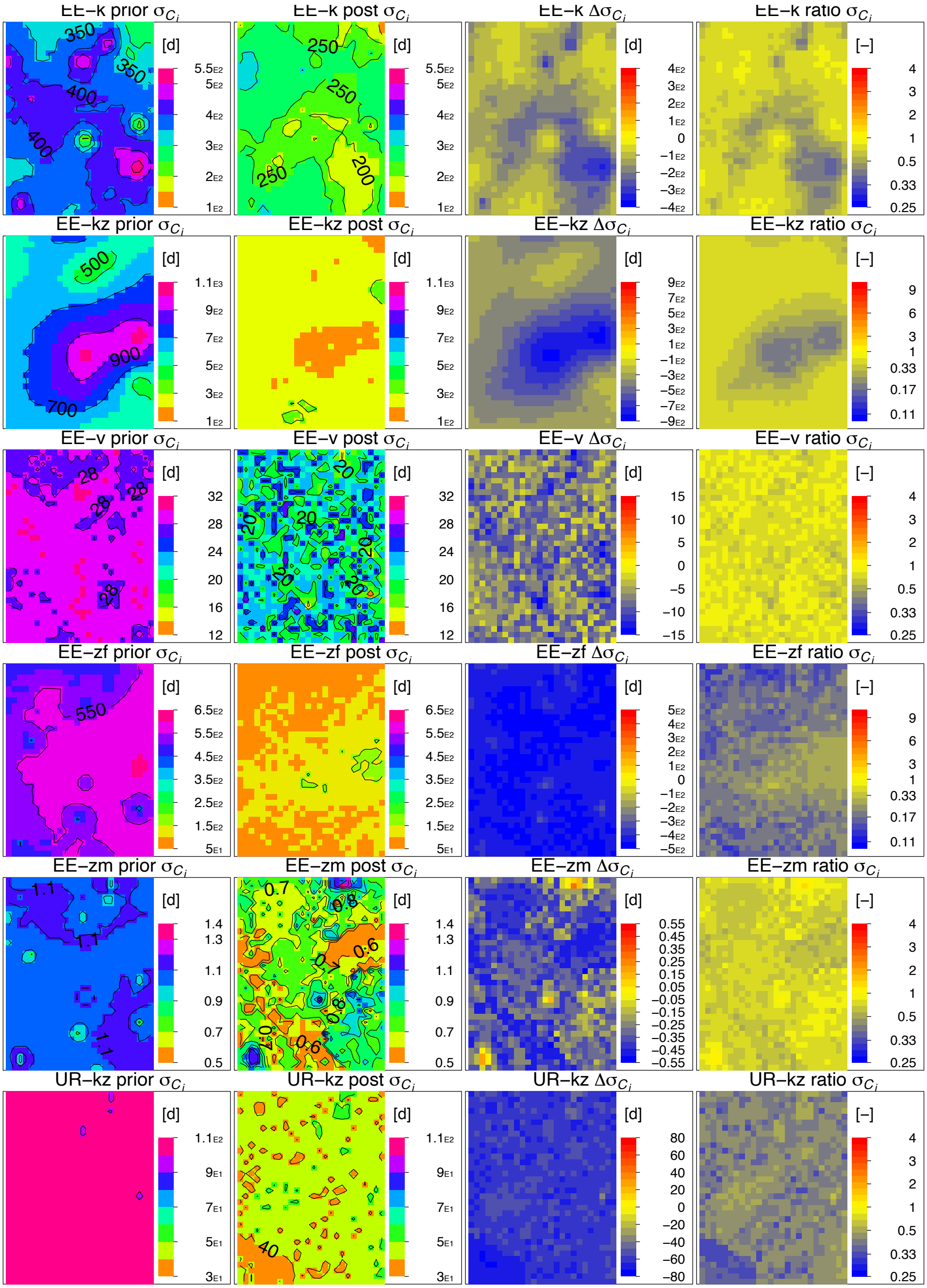

Figure 5.12: *The standard deviation of the vertical hydraulic resistance C_i for each litho-class is depicted per column. From top to bottom row, the prior and posterior mean resistance, and their difference and ratio are shown.*

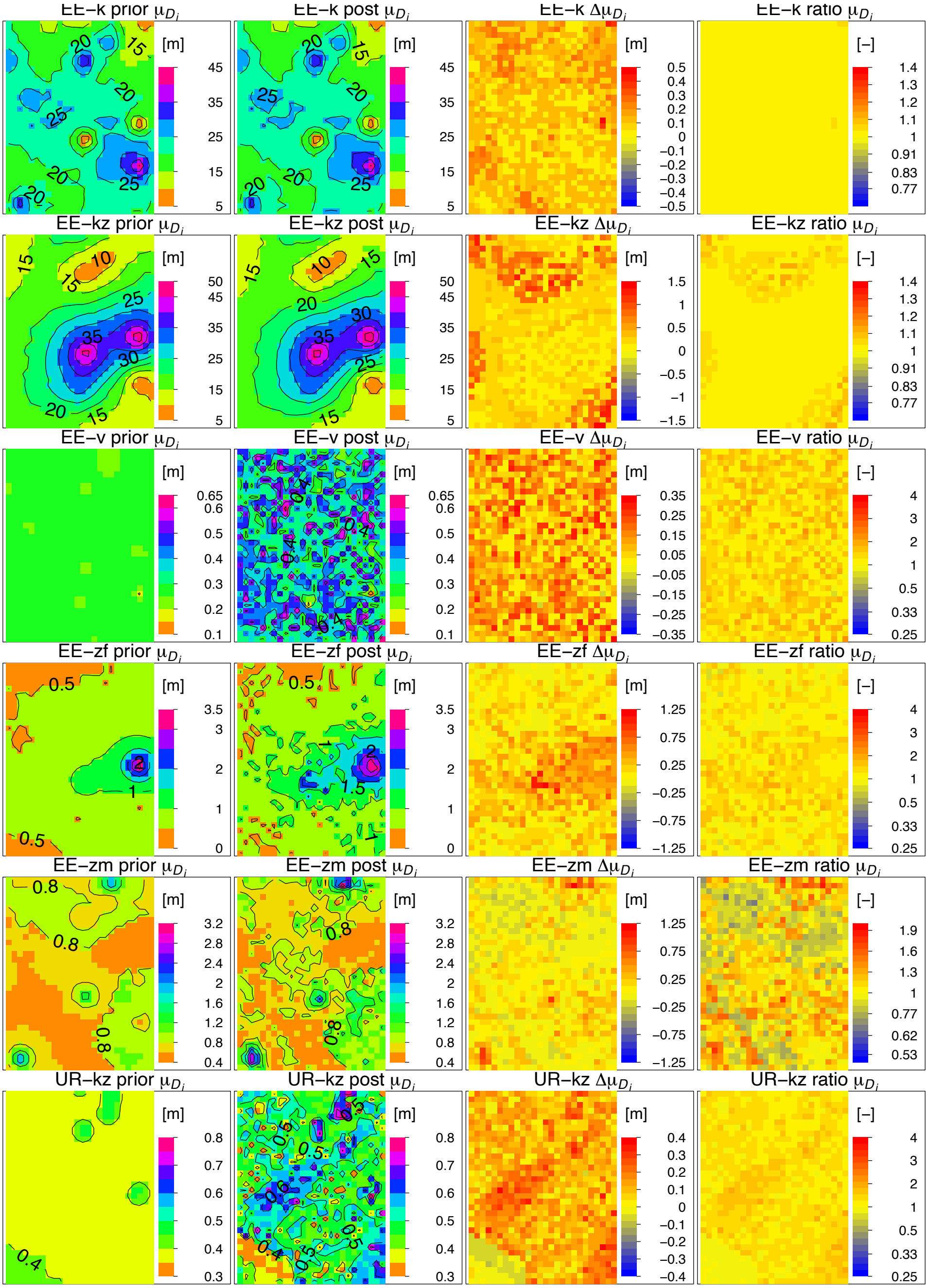

Figure 5.13: *The mean layer thickness D_i of each litho-class is depicted per column. From top to bottom row, the prior and posterior mean value, and their difference and ratio are shown.*

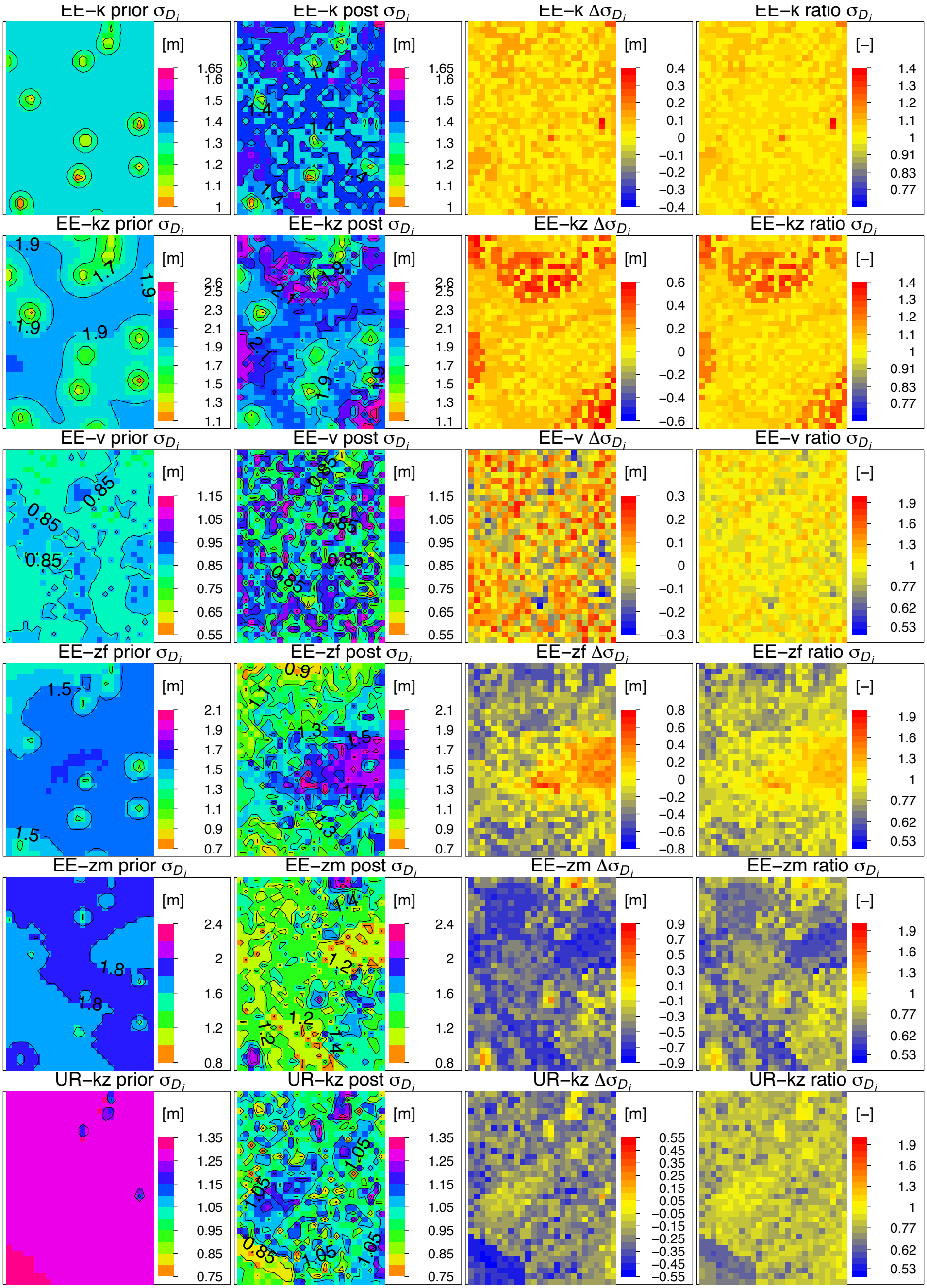
EE−k prior σ_{D_i}
EE−k post σ_{D_i}
EE−k $\Delta\sigma_{D_i}$
EE−k ratio σ_{D_i}
[m] 1.65 1.6 1.5 1.4 1.3 1.2 1.1 1
[m] 1.65 1.6 1.5 1.4 1.3 1.2 1.1 1
[m] 0.4 0.3 0.2 0.1 0 −0.1 −0.2 −0.3 −0.4
[−] 1.4 1.3 1.2 1.1 1 0.91 0.83 0.77
EE−kz prior σ_{D_i}
EE−kz post σ_{D_i}
EE−kz $\Delta\sigma_{D_i}$
EE−kz ratio σ_{D_i}
[m] 2.6 2.5 2.3 2.1 1.9 1.7 1.5 1.3 1.1
[m] 2.6 2.5 2.3 2.1 1.9 1.7 1.5 1.3 1.1
[m] 0.6 0.4 0.2 0 −0.2 −0.4 −0.6
[−] 1.4 1.3 1.2 1.1 1 0.91 0.83 0.77
EE−v prior σ_{D_i}
EE−v post σ_{D_i}
EE−v $\Delta\sigma_{D_i}$
EE−v ratio σ_{D_i}
[m] 1.15 1.05 0.95 0.85 0.75 0.65 0.55
[m] 1.15 1.05 0.95 0.85 0.75 0.65 0.55
[m] 0.3 0.2 0.1 0 −0.1 −0.2 −0.3
[−] 1.9 1.6 1.3 1 0.77 0.62 0.53
EE−zf prior σ_{D_i}
EE−zf post σ_{D_i}
EE−zf $\Delta\sigma_{D_i}$
EE−zf ratio σ_{D_i}
[m] 2.1 1.9 1.7 1.5 1.3 1.1 0.9 0.7
[m] 2.1 1.9 1.7 1.5 1.3 1.1 0.9 0.7
[m] 0.8 0.6 0.4 0.2 0 −0.2 −0.4 −0.6 −0.8
[−] 1.9 1.6 1.3 1 0.77 0.62 0.53
EE−zm prior σ_{D_i}
EE−zm post σ_{D_i}
EE−zm $\Delta\sigma_{D_i}$
EE−zm ratio σ_{D_i}
[m] 2.4 2 1.6 1.2 0.8
[m] 2.4 2 1.6 1.2 0.8
[m] 0.9 0.7 0.5 0.3 0.1 −0.1 −0.3 −0.5 −0.7 −0.9
[−] 1.9 1.6 1.3 1 0.77 0.62 0.53
UR−kz prior σ_{D_i}
UR−kz post σ_{D_i}
UR−kz $\Delta\sigma_{D_i}$
UR−kz ratio σ_{D_i}
[m] 1.35 1.25 1.15 1.05 0.95 0.85 0.75
[m] 1.35 1.25 1.15 1.05 0.95 0.85 0.75
[m] 0.55 0.45 0.35 0.25 0.15 0.05 −0.05 −0.15 −0.25 −0.35 −0.45 −0.55
[−] 1.9 1.6 1.3 1 0.77 0.62 0.53

Figure 5.14: *The standard deviation of the layer thickness D_i of each litho-class is depicted per column. From top to bottom row, the posterior standard deviation, and the difference and ratio of the posterior and prior values are shown. At the top-right corner the standard deviation of the prior distribution is given.*

class separately. Since only one prior distribution is available for the hydraulic conductivity of a litho-class, the maps with the prior means and standard deviations are omitted in these figures. The mean conductivity values of litho-classes EE-k and EE-kz are all decreased, up to a factor of about 0.3. The standard deviations of litho-class EE-k show increase and decrease of the values with a factor between about 1.3 and 0.8. The standard deviations of litho-class EE-kz show a decrease with a factor up to about 0.3. The mean conductivities of the other four litho-classes are either decreased or increased but do not show any lateral variation with the applied legend. The same holds for the standard deviations of these litho-classes.

5.3 Discussion and conclusions

The aim of this chapter was to develop a method for a Bayesian update, and to apply it to hydrogeological model parameters. The method was applied to an aquitard of a groundwater flow model, which in turn is build-up of multiple layers from the REGIS hydrogeological model. These layers are described in terms of litho-classes, where each litho-class has its own hydraulic properties. These properties, layer thickness D and hydraulic conductivity K, are described by probability density functions (PDFs), the prior distributions. As observations, the calibrated aquitard resistances of a groundwater flow model are used.

The probability density functions of D and K are considered to be a prior predictive distribution of the layer thickness and the conductivity, respectively. Thereto, the PDFs are decomposed into stochastic models with a location parameter, a scale parameter and a shape distribution. These three distributions can be arranged in the Bayesian network such that all three are marginal distributions. Herewith, all non-marginal distributions in the whole network become deterministic, which simplifies the calculations. The transformation of the prior predictive distributions into a stochastic model is performed by a Monte Carlo algorithm. This algorithm is able to create marginal distributions with which the prior predictive distributions of D and K are described quite well. The individual marginal distributions have, however, the need for further improvement. This is specifically noticeable by the less smooth shape of the marginal distributions, and the shape of the posterior predictive distributions of the litho-class hydraulic resistance C_i. These last distributions show sometimes multi-modality, where this is not expected given the prior predictive distributions of D and K.

In the described Bayesian network no prerequisites are made for the type of allowed probability density functions. So, advantages of conjugate models, like simple update algorithms, can not be used. Therefore, an iterative algorithm is designed to find the desired likelihood functions. This algorithm performs well, even in finding likelihood peaks in relatively small areas of the range of the marginal distributions.

In the current examples of this study, only the calibrated hydraulic resistance

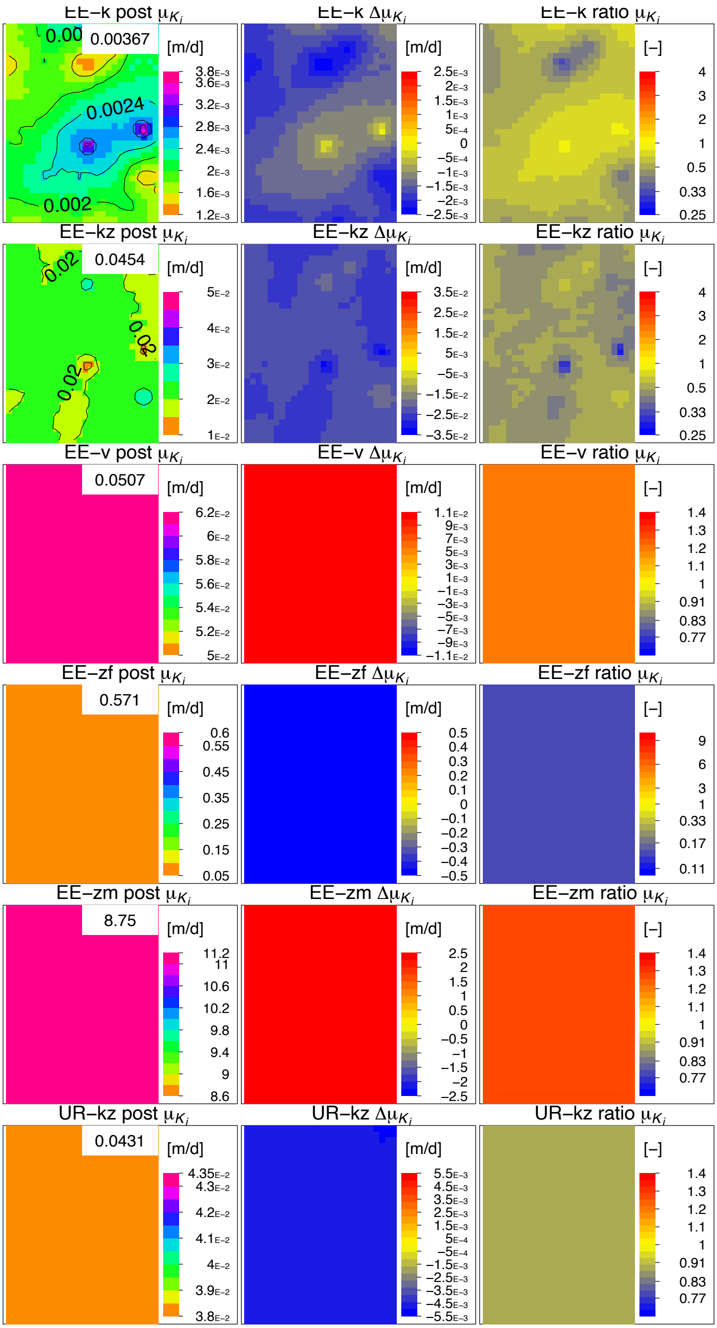

Figure 5.15: *The standard deviation of the vertical hydraulic conductivity K_i of each litho-class is depicted per column. From top to bottom row, the posterior standard deviation, and the difference and ratio of the posterior and prior values are shown. At the top-right corner the standard deviation of the prior distribution is given.*

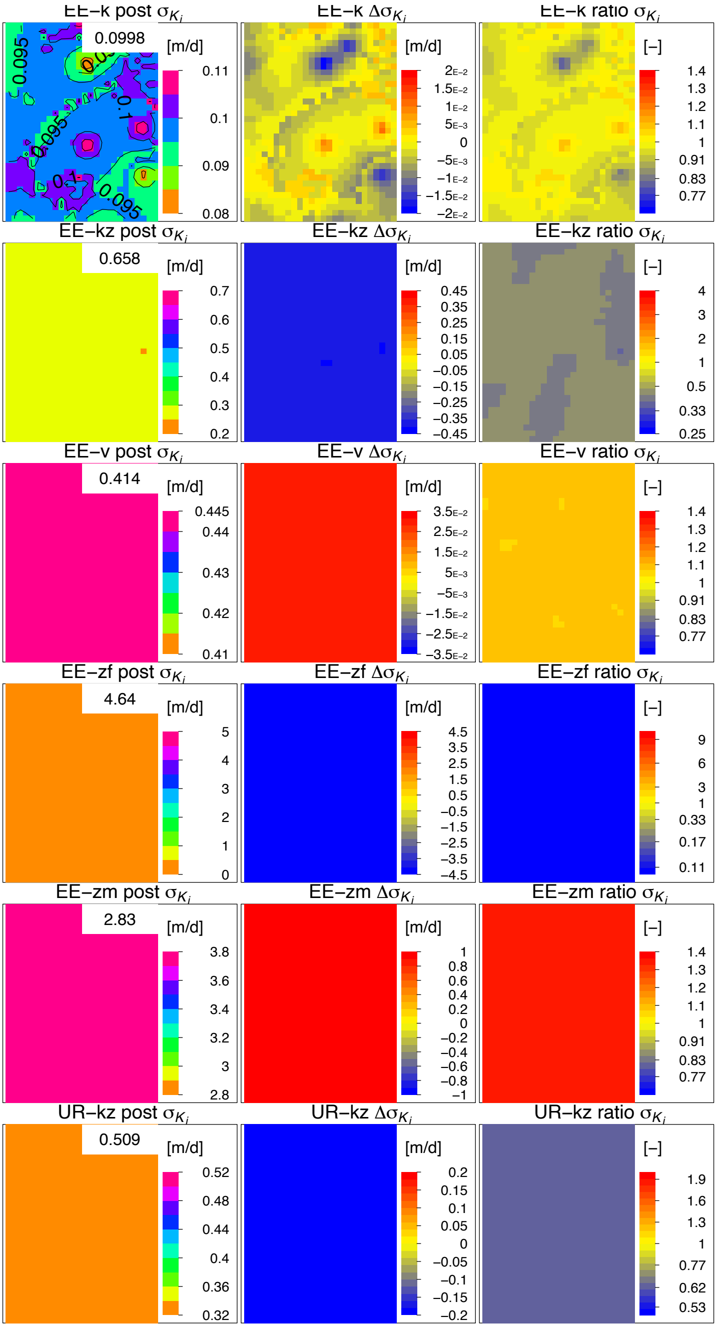

Figure 5.16: *The mean vertical hydraulic conductivity K_i of each litho-class is depicted per column. From top to bottom row, the posterior mean value, and the difference and ratio of the posterior and prior values are shown. At the top-right corner the prior standard deviation is given.*

of the aquitard of one groundwater flow model was used. It is not uncommon that multiple models are available in the same area. If so, the results of all these models can be used as multiple observations in the update process. Since all models are subject to uncertainty, the observations obtained from these models, i.e. the calibrated C values, can be described by a probability density function. Different models may be considered of different reliability, which can be expressed in the PDFs of the observations. However, caution should be taken when using multiple models since their results might be not completely independent.

In the current set-up of the Bayesian network, each grid-cell has its own network with one observation. So the update of the marginal distributions is only locally performed. It makes sense to assume that the probability distribution of the layer thicknesses is local, the interpolated thickness and the uncertainty depends on the interpretation at the borehole locations, and on the distance to the neighboring observations. Nevertheless, a lateral correlation in the layer thickness is very likely otherwise the performed kriging interpolation is not valid. The prior distributions of the hydraulic conductivity consists of one distribution for each litho-class, regardless of the spatial position of the grid cells. It is not unrealistic to assume that these distributions, to some extent, are constant within an area, because they describe a deposit property. To account for a lateral consistency of the distributions, the Bayesian network can be rearranged to have multiple grid cells with each their own thickness marginals, and one common marginal distribution of the conductivity for each litho-class. In this case, multiple observations, i.e. the calibrated parameters of the groundwater flow model, are available to update the conductivity distribution. Since these observations are presumably not uncorrelated, such correlations should be taken into account in the update algorithm.

In one grid cell, the layer thickness and uncertainty for each litho-class is determined (by kriging interpolation). Adding up these individual thicknesses, including their uncertainty, yields a thickness distribution function of the total aquitard thickness. If the total aquitard thickness can also be obtained from another source, an update of the layer thicknesses can be made separately. Another source can be the interpolated total layer thickness, which usually has a lower uncertainty than the summarized thickness of the litho-class thicknesses. With such observations, the proposed calculation of the likelihood with uncertain observations can be applied.

The domains of the initial distributions of the marginal variables are bounded in such a way that, when integrated out, the resulting distribution (D or K) has the same domain as the the prior of that variable. This restriction is applied because the nature of the parameters D and K restricts them to non-negative values. Often, the distribution of a hydraulic conductivity is assumed to be log-normal. Therefore, a log transformation of the distribution of K can be added to the stochastic model to circumvent this problem. It is expected that such a transformation simplifies the creation of the prior marginal distributions. Since a log-transformation is also completely deterministic this can be easily implemented in the presented Bayesian network.

6

Synthesis

SUBSURFACE DATA are widely used in infrastructural projects, groundwater exploitation, environmental assessments and the assessment of subsurface resources (e.g. precious metals and building material). These data involve properties like the type of material (clay, sand, rock) and material properties like hydraulic conductivity and porosity. Such data is usually available as point data, such as borehole descriptions and cone penetration tests, or as line data, like seismic lines. Beyond the observations, no data is available. At the unobserved locations an estimate of subsurface properties can be made by defining a spatial variation model of these properties. Unfortunately, a model is always an interpretation of reality and therefore suffers from imperfections. Moreover, even the observations are to some extent subject to uncertainty. Nevertheless, adding more (reliable) observations will almost always improve the model.

In the Netherlands, the hydrogeological model REGIS is an important source of subsurface data with an emphasis on groundwater flow applications. The REGIS model is a general purpose hydrogeological model. This means that it is not developed for just one application, but it serves as a knowledge base and model for hydrogeological projects. Like all models, also this model has room for improvements. It is possible to collect more hydrogeological point data at unobserved areas, but this is expensive and has only local impact on the model. Instead of collecting more observations of the same properties, it may be beneficial to use available observations of other properties which do have a relation to the subsurface parameters. The REGIS model and parts of it are often applied in groundwater flow models. These groundwater flow models are calibrated using, among others, groundwater head observations. Since these groundwater head observations are not used, or only limited, during building of the REGIS model, this is new data with possibly added value for improvement of REGIS. Herewith we come to the main objective of this thesis:

> *Develop a method or procedure to let the generic hydrogeological model, in our case REGIS, benefit from the improvements of a calibrated groundwater flow model.*

Since all data are subject to uncertainty, a sub-objective is:

> *Develop a method which accounts for uncertain data of all kinds of probability distributions.*

It is likely that, in the Netherlands, multiple calibrated groundwater flow models are available in the same area which may contribute to the improvement of the generic hydrogeological model. Herewith we come to the second sub-objective:

> *Develop a method which can use multiple calibrated groundwater flow models in the same area and with different uncertainty.*

With regard to the main objective of this thesis, two methods have been developed to update the hydrogeological model making use of calibration results of groundwater flow models. The first method (Chapter 3 and Chapter 4) is able to use calibrated data of one groundwater flow model and returns the most likely subsurface parameter values of the litho-layers (thickness and conductivity). The second method (Chapter 5) returns updated probability distributions of the subsurface parameters. Both methods make use of uncertain data of the hydrogeological model, as mentioned in the first sub-objective. The second method is able to handle data of multiple calibrated groundwater flow models, as desired in the second sub-objective. This method is also able to use observations of different uncertainty (also part of the third objective), which is described in Chapter 5. However, this was not implemented yet and therefore no examples could be presented.

6.1 A common thread: uncertainty

A common thread throughout this thesis is uncertainty and how to make calculations with uncertain variables tractable. Analytical and numerical solutions are available to perform such calculations, both with their own advantages and disadvantages. Analytical solutions do in general have a high performance, with regard to calculation time and accuracy, but are not always available. Numerical solutions are more generally applicable but are in general slower and with less accurate results. A hybrid numerical-analytical method can benefit from the advantages of both, and is applied in this thesis.

In Chapter 2, probability density functions (PDFs) were described as piecewise linear functions, and calculations with piecewise linear probability density functions (PL-PDFs) were developed. Herewith, it was possible to perform calculations with (continuous) random variables (RVs) of all kind of distributions with almost no limitations on the shape of these distributions. The piecewise linear description of a PDF is almost always an approximation of the analytical form (if exists) of the distribution, but the description and calculations can be preformed with a negligible loss of accuracy. Even very skewed distributions, like heavy tailed log-normal distributions with a very large standard deviation/mean ratio, can be described adequately. In this context, a negligible loss should be understood as inaccuracies in the calculations which have no impact on the interpretations of the results. As a showcase of the performance of the proposed methods, two examples from the literature are chosen to reproduce their results by using piecewise linear PDFs. These examples are found in Appendix A.2.

The calculations with piecewise linear PDFs were first applied to the kriging interpolation of transmissivities bearing uncertainty of the observations. These observations were obtained from the borehole interpretations where the described litho-class thicknesses and the hydraulic conductivities are supplied with a PDF, describing their respective uncertainties. The PDF of each transmissivity observation, which is the product of the conductivity and the layer thickness, were obtained without a presumed standard or parameterized form of distribution. But by using PL-PDFs, the calculations were not hampered. These transmissivity PDFs could be used to create experimental variograms which clearly showed a nugget effect caused by the uncertainty of the observations. A kriging interpolation was performed on the transmissivity observations which yielded interpolated transmissivity fields honoring the distributions of the observations, without residing to Monte Carlo (MC) solutions.

A second application of the piecewise linear PDFs is finding the most likely parameter values of layer thickness and conductivity in a joint distribution, given an observation of a compound parameter. In Chapter 3, the compound parameter is the aquitard hydraulic resistance, which is the result of an arithmetic combination of the parameters of the hydrogeological model. These uncertain model parameters are described by marginal PDFs of the joint distribution. The most likely marginal parameters are the parameters for which the joint distribution yields its maximum density, given the compound observation. If the joint distribution has an analytical formulation, then it might be possible to find an analytical solution for the above problem. Such a solution has to be derived for each combination of distributions and arithmetic operations, and might even not exist. The choice of piecewise linear PDFs as marginal distributions yielded a tractable solution.

As a last application of PL-PDFs, an aquitard, consisting of multiple litho-layers, was cast in a Bayesian network (BN). Again, the calibrated aquitard resistances of a groundwater flow model served as observations. The deterministic relations between the parameters could be defined using PL-PDFs. The aim of the Bayesian network is to update the prior uncertainty of the parameters of the layer thicknesses and hydraulic conductivities, given the observations. Creating a stochastic model for these parameters had quite some freedom using PL-PDFs. It was shown that the distinction between a stochastic node and a deterministic node in a BN is not necessarily very strict. These nodes could often be converted from one type into the other. The likelihood functions, needed to find the posterior marginal distributions, were also described as piecewise linear functions. In the literature, this description of the likelihood function is often mentioned as problematic in finding an adequate discretization of the function, especially when it has large tails or when a large amount of the integral resides in a small area of its domain. An algorithm was developed to circumvent this discretization problem.

No doubt, analytical solutions, when available, have a much smaller calculation time compared to calculations with piecewise linear PDFs. But the advantage of PL-

PDFs over analytical solutions is that they are available for a wider range of PDFs. It is not necessary to reside to distributions which have an analytical solution. And for new problems, no additional tedious analytical derivations are needed.

6.2 Update methods

The main objective of this research was to develop a method with which extra data can be used to improve the hydrogeological model REGIS, in particular data from calibrated groundwater flow models. The REGIS model is horizontally discretized in grid cells, and in vertical direction in layers called hydrogeological units. Each hydrogeological unit can consist of multiple depositional classes, so called litho-classes. To each combination of grid cell, hydrogeological unit and litho-class, properties like conductivity and layer thickness are assigned. These properties are usually the litho-class properties, but are assigned within the context of their location in the subsurface denoted by grid cell, hydrogeological unit combination. At each location these properties may differ.

The hydrogeological units of the REGIS model are defined by their top and bottom, and by the hydraulic conductivity. For aquifers the transmissivity is given and for aquitards the vertical resistance. As said, a hydrogeological unit may consist of multiple litho-classes and its properties is therefore an aggregation of the properties of these litho-classes. Since in REGIS the hydraulic conductivity is defined at the level of a litho-class, it is necessary to know the contribution of each litho-class to the properties of the hydrogeological units at each grid cell. Therefore, the interpolation of the litho-class thicknesses is performed in this study. This interpolation started from the borehole descriptions and interpretations. Borehole descriptions usually contain interval depths which are assumed to be deterministic. Nevertheless, these thicknesses are subject to uncertainty. Therefore, a method is proposed to assign a PDF to each litho-layer thickness which depends on the round-off value of the thickness. Subsequently, these PDFs are used in the interpolation of the thicknesses.

Two update methods have been proposed, both with different qualities. The first method describes, at each grid cell, the joint distribution of the layer thickness and the hydraulic conductivity of all litho-classes. In the presented examples, 7 litho-classes are recognized, which yields a 14-dimensional joint distribution. The 14 marginal distributions are the PDFs of the layer thickness and the hydraulic conductance of each litho-class. In this distribution, the most likely combination of marginal values was determined. The most likely value is defined as the marginal values combination with the highest joint probability density, which is the mode of the joint PDF. Without any constraints, each marginal value would get the mode of its marginal distribution. But the calibrated value, an aquitard resistance of the groundwater flow model, is used as a constraint, which means that the combination of the marginal values must exactly yield the calibrated value. In general, infinite combinations of marginal values will meet the constraint to form the calibrated value but usually only a limited number of combinations, or even just one, will yield a maximum probability density. An algorithm was developed to find this constrained mode in a multi-dimensional joint distribution. The result is one value for each marginal variable, so no information about the uncertainty is left.

The second method uses the same joint distributions as the first one, but these distributions are now cast in a Bayesian network (BN). The difference between the joint distributions is that now the distributions of the layer thickness and the conductivity are converted into stochastic models. Also in this method, the calibrated aquitard resistance is used as an observation to update the prior distributions. With these observations, a Bayesian update of the marginal distributions is performed.

The use of a Bayesian network has some advantages over the first method (finding the most likely value). Firstly, after performing an updating the parameters of interest are still described by probability distributions. So knowledge about uncertainty is retained. Secondly, multiple observations can be used for an update. The Bayesian update makes use of a likelihood function which always can accept multiple observations. Thirdly, the Bayesian update can make use of uncertain observations. In the context of this thesis, this is especially useful when multiple calibrated groundwater flow models are available in the same area but with a different reliability. This option is described in Chapter 5, although not yet applied. The first method has also advantages over the second one. The method is easier to implement and the calculation times are much lower.

The parameterization of the hydraulic conductivity in REGIS depends only on the assigned litho-class, but the same litho-class may have different conductivity properties at different locations in the Netherlands. For instance, if the deposits of the same litho-class have been buried in one area and not in another, then the conductivities are expected to be lower in the first area. These types of differences are not accounted for in the current version of the REGIS model. If this difference is reflected in the calibrated data, then it could be recognized in the updated values. This test is performed for the same aquitard in two distinct areas, with the first update method applied. Instead of just looking at the updated values in a cell or in an area, the update patterns of different areas were compared. As a measure for comparison a cumulative frequency distribution (CFD) of the most likely conductivity values was created per litho-class for each area. The CFDs appeared to be clearly different, which strongly suggest different hydraulic properties between these areas.

6.3 Some room for improvement

Several methods have been presented to improve the parameterization of a generic hydrogeological model by using information of calibrated groundwater flow models. All given examples were applied to either an aquifer or an aquitard, but not to both. The methods are nevertheless not restricted to a single layer type. The story does not end here, as several improvements are recognized en described hereafter.

In all applications in this thesis, uncertain values of hydraulic conductivity and layer thickness of litho-layers are used. The PDFs of the conductivity are used as defined in the REGIS system. This means, the same distribution for every grid-cell containing the same litho-class deposits. The litho-layer thickness for each grid-cell is obtained by interpolation and has a dedicated distribution for each grid-cell for each litho-class. At any location in the model, the summation of all layer thicknesses yields a PDF of the total layer thickness. This total layer thickness might also be available from any other source with a lower uncertainty. In REGIS, the top and bottom of all defined hydrogeological units are determined. The aquitards and aquifers of a groundwater flow model are often an aggregation of multiple hydrogeological units. The top and bottom of this aggregated layer can be used as another source of the layer thickness. In the proposed update method with the Bayesian network (Chapter 5), first a posterior layer thickness can be determined using a BN containing only layer thicknesses. Thereafter, the newly obtained layer thickness distributions can be used in the update with the calibrated parameters of the groundwater flow model.

The proposed Bayesian update yields for each location (grid cell) a posterior distribution of the hydraulic conductivity. The prior distribution is just one PDF for all locations. The truth might be somewhere in-between, a PDF for each homogeneous considered area. The BN can be reconfigured to support this. An advantage is that more data is used to update one distribution.

To make a Bayesian update of the hydrogeological model parameters (layer thickness and conductivity) attainable, a stochastic model of the RVs must be defined. In Chapter 5, this is achieved by the decomposition of a PDF into a location-scale-shape model, with, in the context of a BN, only the leaf nodes being stochastic. All other (internal) nodes are deterministic. The decomposition is performed through a MC method, which did not always yield a satisfactory result. This poor result was caused by used distributions whose shapes were hard to catch in a location-scale-shape model. Standard stochastic models, other than location-scale-shape models, are available which might describe the distributions more adequately. Usually, these models are also described with stochastic internal nodes. This restricts the shape of the distribution to some standard parameterized form. With only leaf nodes being stochastic, the freedom of shape is larger. As shown in Chapter 5, the difference between stochastic and deterministic nodes is not always very strict. So a decomposition into some standard or conjugate model, instead of trying to fit every PDF into a location-scale-shape variant, could be possible. It may sound odd in the context of this thesis to promote the use of standard stochastic models, since the use of PL-PDFs was eulogized for its independence of such models. Nevertheless, it can be very useful when a BN has combinations of distributions for which an analytical solution of a likelihood function does not exist, or the location-scale-shape decomposition appears not very adequate. Thereby, one assumption is that the distributions of the hydrogeological model parameters are known and represent the prior predictive distributions of these. Therefore, integrating out the marginal variables of the stochastic model of these parameters should exactly yield the prior predictive distribution. So starting with some standard model, convert eventually internal stochastic nodes to deterministic, and then reshape the marginal distributions to meet the requirements of the prior predictive distribution. This might be a better starting point for the stochastic model but keeps the flexibility of the usage of PL-PDFs. As a simple example, a decomposition of an RV with a PDF with a shape close to a log-normal distribution can yield a quite cumbersome result when forced into a location-scale-shape stochastic model, as seen in the examples. An additional deterministic step could be to first take the logarithm of the RV, which would yield a function close to a Gaussian distribution, and decompose this last distribution into a location-scale-shape model.

In Chapter 5, the posterior distributions of the BN are obtained assuming the observations have no uncertainty. This is not a realistic assumption but to some extent defensible. However, if an uncertainty of the observations can be quantified, it would be better to incorporate that information in the update procedure. Moreover, if multiple observations are available with different reliability, it is necessary to account for the uncertainty. A method has been described to apply the uncertain observations to the calculations of the likelihood function. This extension to the Bayesian update has not been applied yet in the given examples, but should be implemented in future work. Especially when multiple realizations of calibrated groundwater flow models are available with different uncertainty.

One important source of uncertainty has not been addressed: the assignment of a litho-class to an interval in the borehole descriptions. This classification depends highly on the quality of the description of the depositional material, which in addition depends highly on the drilling method, and, in case a classification is made in the field, on the experience of the field geologist. Also, from the description it may be hard to distinguish to which class the interval should be assigned. The assignment of the wrong litho-class to an interval does have consequences for the assumed hydraulic properties of that interval. Since a classification system is used, the assignment of the one or the other class may yield a conductivity which differs one or two orders in magnitude. This is an important subject for future research.

A

Calculations with piecewise linear PDFs

A.1 Elementary operations

This appendix describes the derivation of four elementary binary operations ($+ - \times /$) performed on piecewise linear probability density functions (PDFs) as proposed in Chapter 2. At the end of this appendix, a performance example is given where the piecewise linear calculations are compared to numerical examples from the literature.

A.1.1 Probability distributions of binary operations

Let X and Y be independent random variables (RVs) and Z be the result of a binary operation on X and Y. The general formulation of the cumulative distribution function (CDF) of Z can be written as *Papoulis* [1991, p. 132 ff]

$$F_z(z) = \iint f_x(x) f_y(y) \, \mathrm{d}x \, \mathrm{d}y, \tag{A.1}$$

where $f_x(\cdot)$ and $f_y(\cdot)$ are the PDFs of X and Y, respectively. These PDFs are linear functions at each bin of the piecewise linear PDFs and are, for bin i and bin j, defined as

$$f_{x,i}(x) = p_{x_i} + r_{x_i}(x - x_i) \tag{A.2}$$

$$f_{y,j}(y) = p_{y_j} + r_{y_j}(y - y_j), \tag{A.3}$$

where p_{x_i} and p_{y_j} are the probability densities at the values x_i and y_j, respectively. The slopes of these functions are defined as $r_{x_i} = (p_{x_{i+1}} - p_{x_i})/(x_{i+1} - x_i)$ and $r_{y_j} = (p_{y_{j+1}} - p_{y_j})/(y_{j+1} - y_j)$. For convenience, the next variables are defined

$$\begin{aligned} p_{0,x_i} &= f_{x,i}(0) = p_{x_i} - r_{x_i} x_i \\ p_{0,y_j} &= f_{y,j}(0) = p_{y_j} - r_{y_j} y_j. \end{aligned} \tag{A.4}$$

Since the functions $f_{x,i}(\cdot)$ and $f_{y,j}(\cdot)$ are only continuously within a bin, Equation (A.1) has to be defined for each joint bin as

$$F_{z,ij}(z) = \iint f_{x,i}(x) f_{y,j}(y) \, \mathrm{d}x \, \mathrm{d}y, \tag{A.5}$$

Furthermore, the integration area of a joint bin is split up into four sub-areas, shown in Figure A.1. As can be seen, the integration boundaries $x_{l,i}$, $x_{u,i}$, $y_{l,j}$ and $y_{u,j}$ depend on the intersection of the line $z = g(x, y)$ with the lines $x = x_i$, $x = x_{i+1}$, $y = y_j$ and $y = y_{j+1}$. The function $g(x, y)$ represents a binary operation.

The line $z = g(x, y)$ for a particular value of z will not intersect all joint bins. Therefore z_{ij} is defined as z but limited to the minimum and maximum value of z for which $g(x, y)$ intersects joint bin (i, j).

The probabilities of the rectangle sub-areas a, b and c can be easily defined by the product of their marginal probabilities

$$\begin{aligned} F_{z,ij,a}(z) &= \Pr\{x_i < X \le x_{l,i}\} \Pr\{y_j < Y \le y_{l,j}\} \\ F_{z,ij,b}(z) &= \Pr\{x_{l,i} < X \le x_{u,i}\} \Pr\{y_j < Y \le y_{l,j}\} \\ F_{z,ij,c}(z) &= \Pr\{x_i < X \le x_{l,i}\} \Pr\{y_{l,j} < Y \le y_{u,j}\}. \end{aligned} \tag{A.6}$$

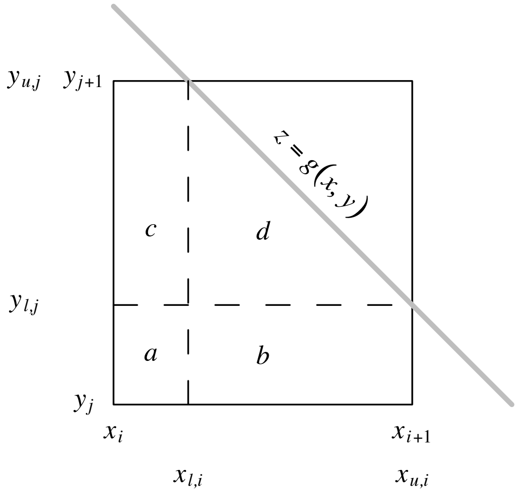

Figure A.1: *Integration boundaries of the piecewise analytical* CDF. *Shown is the dependence of the integration boundaries on the position of the line z in the box of the joint bin (i, j). The function $g(\cdot)$ denotes any binary operation.*

These three functions hold for the example in Figure A.1, the boundaries may be different for other operations. The function for sub-area d ($F_{z,ij,d}(z)$) is described by Equation (A.5) and is derived for each binary operation separately in the next sections. The probability of $Z < z$ for bin (i, j) for a given value of z is defined as

$$F_{z,ij}(z) = F_{z,ij,a}(z) + F_{z,ij,b}(z) + F_{z,ij,c}(z) + F_{z,ij,d}(z). \tag{A.7}$$

To obtain the cumulative probability for a particular value of Z, a summation of the probabilities of all joint bins has to be performed

$$F_z(z) = \sum_{j=1}^{n_y} \sum_{i=1}^{n_x} F_{z,ij}(z), \tag{A.8}$$

where n_x and n_y are the numbers of bins of X and Y, respectively.

Subsequently, the first derivative of $F_z(z)$ with respect to z is the corresponding PDF. The PDF is calculated as the derivative of $F_{z,ij,d}(z)$ only, the probabilities of the areas a, b and c are constant values in this context.

Summation

Let $Z = X + Y$. The integration boundaries for joint bin (i, j) are defined as

$$\begin{aligned} y_{u,j} &= \max(y_j, \min(y_{j+1}, z - x_i)) \\ y_{l,j} &= \max(y_j, \min(y_{j+1}, z - x_{i+1})) \\ x_{u,i} &= \max(x_i, \min(x_{i+1}, z - y_{l,j})) \\ x_{l,i} &= \max(x_i, \min(x_{i+1}, z - y_{u,j})) \\ z_{ij} &= x_{u,i} + y_{l,j} = x_{l,i} + y_{u,j}. \end{aligned} \tag{A.9}$$

Equation (A.5) for sub-area d can be written as

$$\begin{aligned} F_{z,ij,d}(z) &= \int_{y=y_{l,j}}^{y_{u,j}} \int_{x=x_{l,i}}^{z_{ij}-y} f_{x,i}(x) f_{y,j}(y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{y=y_{l,j}}^{y_{u,j}} \int_{x=x_{l,i}}^{z_{ij}-y} \left[p_{0,x_i} + r_{x_i}x\right] f_{y,j}(y)\,\mathrm{d}x\,\mathrm{d}y. \end{aligned} \tag{A.10}$$

Integration with respect to x yields

$$F_{z,ij,d}(z) = \int_{y=y_{l,j}}^{y_{u,j}} \left[\,\left(p_{0,x_i}x + \tfrac{1}{2}r_{x_i}x^2\right)\,\right]_{x=x_{l,i}}^{z_{ij}-y} f_{y,j}(y)\,\mathrm{d}y. \tag{A.11}$$

Inserting integration boundaries yields

$$F_{z,ij,d}(z) = \int_{y=y_{l,j}}^{y_{u,j}} \left[p_{0,x_i}(z_{ij} - y - x_{l,i}) + \tfrac{1}{2}r_{x_i}\left((z_{ij}-y)^2 - x_{l,i}^2\right)\,\right] f_{y,j}(y)\,\mathrm{d}y. \tag{A.12}$$

Substituting $((z_{ij}-y)^2 - x_{l,i}^2)$ by $((z_{ij}-y-x_{l,i})^2 + 2x_{l,i}(z_{ij}-y-x_{l,i}))$, $p_{x_{l,i}} = f_{x,i}(x_{l,i})$ and $z_{ij} - x_{l,i} = y_{u,j}$ yields

$$F_{z,ij,d}(z) = \int_{y=y_{l,j}}^{y_{u,j}} \left[p_{x_{l,i}}(y_{u,j}-y) + \tfrac{1}{2}r_{x_i}(y_{u,j}-y)^2\right]\left[p_{0,y_j} + r_{y_j}y\right]\mathrm{d}y. \tag{A.13}$$

Substituting $r_{y_j}y = -r_{y_j}(y_{u,j}-y) + r_{y_j}y_{u,j}$, and $p_{y_{u,j}} = f_{y,j}(y_{u,j})$ yields

$$F_{z,ij,d}(z) = \int_{y=y_{l,j}}^{y_{u,j}} \left[p_{x_{l,i}}(y_{u,j}-y) + \tfrac{1}{2}r_{x_i}(y_{u,j}-y)^2\right]\left[p_{y_{u,j}} - r_{y_j}(y_{u,j}-y)\right]\mathrm{d}y. \tag{A.14}$$

Integration with respect to y yields

$$\begin{aligned} F_{z,ij,d}(z) = \Big[\, & p_{x_{l,i}}p_{y_{u,j}}(-\tfrac{1}{2})(y_{u,j}-y)^2 - p_{x_{l,i}}r_{y_j}(-\tfrac{1}{3})(y_{u,j}-y)^3 \\ & + \tfrac{1}{2}r_{x_i}p_{y_{u,j}}(-\tfrac{1}{3})(y_{u,j}-y)^3 - \tfrac{1}{2}r_{x_i}r_{y_j}(-\tfrac{1}{4})(y_{u,j}-y)^4\Big]_{y=y_{l,j}}^{y_{u,j}}. \end{aligned} \tag{A.15}$$

Inserting integration boundaries yields

$$\begin{aligned} F_{z,ij,d}(z) = \; & \tfrac{1}{2}p_{x_{l,i}}p_{y_{u,j}}(y_{u,j}-y_{l,j})^2 - \tfrac{1}{3}p_{x_{l,i}}r_{y_j}(y_{u,j}-y_{l,j})^3 \\ & + \tfrac{1}{6}r_{x_i}p_{y_{u,j}}(y_{u,j}-y_{l,j})^3 - \tfrac{1}{8}r_{x_i}r_{y_j}(y_{u,j}-y_{l,j})^4. \end{aligned} \tag{A.16}$$

The first derivative of Equation (A.16) with respect to z_{ij} is its corresponding PDF. The variables dependent on z_{ij} are $y_{u,j} = z_{ij} - x_{l,i}$, $x_{u,i} = z_{ij} - y_{l,j}$, and $p_{y_{u,j}} = f_{y,j}(y_{u,j}) = p_{0,y_j} + r_{y_j}(z_{ij} - x_{l,i})$. So the derivative writes

$$\begin{aligned} f_{z,ij,d}(z) = \; & \tfrac{1}{2}p_{x_{l,i}}r_{y_j}(z_{ij}-x_{l,i}-y_{l,j})^2 \\ & + \tfrac{1}{2}p_{x_{l,i}}p_{y_{u,j}}2(z_{ij}-x_{l,i}-y_{l,j}) \\ & - \tfrac{1}{3}p_{x_{l,i}}r_{y_j}3(z_{ij}-x_{l,i}-y_{l,j})^2 \\ & + \tfrac{1}{6}r_{x_i}r_{y_j}(z_{ij}-x_{l,i}-y_{l,j})^3 \\ & + \tfrac{1}{6}r_{x_i}p_{y_{u,j}}3(z_{ij}-x_{l,i}-y_{l,j})^2 \\ & - \tfrac{1}{8}r_{x_i}r_{y_j}4(z_{ij}-x_{l,i}-y_{l,j})^3, \end{aligned} \tag{A.17}$$

and can be rewritten as

$$\begin{aligned} f_{z,ij,d}(z) = \quad & p_{x_{l,i}} p_{y_{u,j}} (y_{u,j} - y_{l,j}) \\ & -\tfrac{1}{2} p_{x_{l,i}} r_{y_j} (y_{u,j} - y_{l,j})^2 \\ & +\tfrac{1}{2} r_{x_i} p_{y_{u,j}} (y_{u,j} - y_{l,j})^2 \\ & -\tfrac{1}{3} r_{x_i} r_{y_j} (y_{u,j} - y_{l,j})^3. \end{aligned} \tag{A.18}$$

Subtraction

Let $Z = X - Y$. The integration boundaries for joint bin (i, j) are defined as

$$\begin{aligned} y_{u,j} &= \max(y_j, \min(y_{j+1}, x_{i+1} - z)) \\ y_{l,j} &= \max(y_j, \min(y_{j+1}, x_i - z)) \\ x_{u,i} &= \max(x_i, \min(x_{i+1}, z + y_{u,j})) \\ x_{l,i} &= \max(x_i, \min(x_{i+1}, z + y_{l,j})) \\ z_{ij} &= x_{u,i} - y_{u,j} = x_{l,i} - y_{l,j}. \end{aligned} \tag{A.19}$$

Equation (A.5) for sub-area d can be written as

$$F_{z,ij,d}(z) = \int_{y=y_{l,j}}^{y_{u,j}} \int_{x=x_{l,i}}^{z_{ij}+y} f_{x,i}(x) f_{y,j}(y) \, \mathrm{d}x \, \mathrm{d}y. \tag{A.20}$$

Integration with respect to x yields

$$\begin{aligned} F_{z,ij,d}(z) = \int_{y=y_{l,j}}^{y_{u,j}} & \left[p_{0,x_i}(z_{ij} + y - x_{l,i}) + \tfrac{1}{2} r_{x_i} \left((z_{ij} + y)^2 - x_{l,i}^2 \right) \right] \\ & f_{y,j}(y) \, \mathrm{d}y. \end{aligned} \tag{A.21}$$

Substituting $((z_{ij}+y)^2 - x_{l,i}^2)$ by $((z_{ij}+y-x_{l,i})^2 + 2x_{l,i}(z_{ij}+y-x_{l,i}))$, $p_{x_{l,i}} = f_{x,i}(x_{l,i})$ and $z_{ij} - x_{l,i} = -y_{l,j}$ yields

$$\begin{aligned} F_{z,ij,d}(z) = \int_{y=y_{l,j}}^{y_{u,j}} & \left[p_{x_{l,i}}(y - y_{l,j}) + \tfrac{1}{2} r_{x_i} (y - y_{l,j})^2 \right] \\ & \left[(p_{y_j} - r_{y_j} y_j) + r_{y_j} y \right] \mathrm{d}y. \end{aligned} \tag{A.22}$$

Substituting $r_{y_j} y = r_{y_j}(y - y_{l,j}) + r_{y_j} y_{l,j}$, and $p_{y_{l,j}} = f_{y,j}(y_{l,j})$ yields

$$\begin{aligned} F_{z,ij,d}(z) = \int_{y=y_{l,j}}^{y_{u,j}} & \left[p_{x_{l,i}}(y - y_{l,j}) + \tfrac{1}{2} r_{x_i} (y - y_{l,j})^2 \right] \\ & \left[p_{y_{l,j}} + r_{y_j}(y - y_{l,j}) \right] \mathrm{d}y. \end{aligned} \tag{A.23}$$

Integration with respect to y yields

$$\begin{aligned} F_{z,ij,d}(z) = \Big[& \, p_{x_{l,i}} p_{y_{l,j}} (\tfrac{1}{2})(y - y_{l,j})^2 + p_{x_{l,i}} r_{y_j} (\tfrac{1}{3})(y - y_{l,j})^3 \\ & + \tfrac{1}{2} r_{x_i} p_{y_{l,j}} (\tfrac{1}{3})(y - y_{l,j})^3 + \tfrac{1}{2} r_{x_i} r_{y_j} (\tfrac{1}{4})(y - y_{l,j})^4 \Big]_{y=y_{l,j}}^{y_{u,j}}. \end{aligned} \tag{A.24}$$

Inserting integration boundaries yields

$$\begin{aligned} F_{z,ij,d}(z) = \quad & \tfrac{1}{2} p_{x_{l,i}} p_{y_{l,j}} (y_{u,j} - y_{l,j})^2 + \tfrac{1}{3} p_{x_{l,i}} r_{y_j} (y_{u,j} - y_{l,j})^3 \\ & + \tfrac{1}{6} r_{x_i} p_{y_{l,j}} (y_{u,j} - y_{l,j})^3 + \tfrac{1}{8} r_{x_i} r_{y_j} (y_{u,j} - y_{l,j})^4. \end{aligned} \tag{A.25}$$

The first derivative of Equation (A.25) with respect to z is its corresponding PDF. The variables dependent on z_{ij} are $x_{u,i} = z_{ij} + y_{u,j}$, $y_{l,j} = x_{l,i} - z_{ij}$ and $p_{y_{l,j}} = f_{y,j}(y_{l,j}) = p_{0,y_j} + r_{y_j}(x_{l,i} - z_{ij})$. So the derivative writes

$$\begin{aligned} f_{z,ij,d}(z) = \; & -\tfrac{1}{2} p_{x_{l,i}} r_{y_j} (y_{u,j} - y_{l,j})^2 \\ & +\tfrac{2}{2} p_{x_{l,i}} p_{y_{l,j}} (y_{u,j} - y_{l,j}) \\ & +\tfrac{3}{3} p_{x_{l,i}} r_{y_j} (y_{u,j} - y_{l,j})^2 \\ & -\tfrac{1}{6} r_{x_i} r_{y_j} (y_{u,j} - y_{l,j})^3 \\ & +\tfrac{3}{6} r_{x_i} p_{y_{l,j}} (y_{u,j} - y_{l,j})^2 \\ & +\tfrac{4}{8} r_{x_i} r_{y_j} (y_{u,j} - y_{l,j})^3, \end{aligned} \tag{A.26}$$

and can be rewritten as

$$\begin{aligned} f_{z,ij,d}(z) = \; & \quad p_{x_{l,i}} p_{y_{l,j}} (y_{u,j} - y_{l,j}) \\ & +\tfrac{1}{2} p_{x_{l,i}} r_{y_j} (y_{u,j} - y_{l,j})^2 \\ & +\tfrac{1}{2} r_{x_i} p_{y_{l,j}} (y_{u,j} - y_{l,j})^2 \\ & +\tfrac{1}{3} r_{x_i} r_{y_j} (y_{u,j} - y_{l,j})^3. \end{aligned} \tag{A.27}$$

Multiplication

Let $Z = XY$. For multiplication integration of probability for joint bins has to be performed separately for each quadrant, as can be seen in Figure 2.3. In this section, integration for quadrant 1 ($z \in \langle 0, \infty \rangle$) is derived. The integration boundaries for joint bin (i, j) are defined as

$$\begin{aligned} x_{l,i} &= \max(x_i, \min(x_{i+1}, z/y_{j+1})) \\ y_{l,j} &= \max(y_j, \min(y_{j+1}, z/x_{i+1})) \\ x_{u,i} &= \max(x_i, \min(x_{i+1}, z/y_{l,j})) \\ y_{u,j} &= \max(y_j, \min(y_{j+1}, z/x_{l,i})) \\ z_{ij} &= x_{u,i} y_{l,j} = x_{l,i} y_{u,j}. \end{aligned} \tag{A.28}$$

Equation (A.5) for sub-area d can be written as

$$F_{z,ij,d}(z) = \int_{y=y_{l,j}}^{y_{u,j}} \int_{x=x_{l,i}}^{z_{ij}/y} f_{x,i}(x) f_{y,j}(y) \, \mathrm{d}x \, \mathrm{d}y. \tag{A.29}$$

Integration with respect to x yields

$$\begin{aligned} F_{z,ij,d}(z) = \int_{y=y_{l,j}}^{y_{u,j}} & \left[p_{0,x_i} (z_{ij}/y - x_{l,i}) + \tfrac{1}{2} r_{x_i} \left((z_{ij}/y)^2 - x_{l,i}^2 \right) \right] \\ & f_{y,j}(y) \, \mathrm{d}y. \end{aligned} \tag{A.30}$$

Integration with respect to y yields

$$\begin{aligned} F_{z,ij,d}(z) = \; & \Big[\, p_{0,x_i} p_{0,y_j} (z_{ij} \ln|y| - x_{l,i} y) + p_{0,x_i} r_{y_j} (z_{ij} y - \tfrac{1}{2} x_{l,i} y^2) \\ & + \tfrac{1}{2} r_{x_i} p_{0,y_j} (-z_{ij}^2/y - x_{l,i}^2 y) + \tfrac{1}{2} r_{x_i} r_{y_j} (z_{ij}^2 \ln|y| - \tfrac{1}{2} x_{l,i}^2 y^2) \Big]_{y=y_{l,j}}^{y_{u,j}}. \end{aligned} \tag{A.31}$$

Inserting integration boundaries yields

$$\begin{aligned} F_{z,ij,d}(z) = \quad & p_{0,x_i} p_{0,y_j} (z_{ij} \ln |y_{u,j}/y_{l,j}| - x_{l,i}(y_{u,j} - y_{l,j})) \\ & + p_{0,x_i} r_{y_j} (z_{ij}(y_{u,j} - y_{l,j}) - \tfrac{1}{2} x_{l,i}(y_{u,j}^2 - y_{l,j}^2)) \\ & + \tfrac{1}{2} r_{x_i} p_{0,y_j} (-z_{ij}^2 (y_{u,j}^{-1} - y_{l,j}^{-1}) - x_{l,i}^2 (y_{u,j} - y_{l,j})) \\ & + \tfrac{1}{2} r_{x_i} r_{y_j} (z_{ij}^2 \ln |y_{u,j}/y_{l,j}| - \tfrac{1}{2} x_{l,i}^2 (y_{u,j}^2 - y_{l,j}^2)). \end{aligned} \tag{A.32}$$

The first derivative of Equation (A.32) with respect to z_{ij} is its corresponding PDF. The variables dependent on z_{ij} are $x_{u,i} = z_{ij}/y_{l,j}$, $y_{u,j} = z_{ij}/x_{l,i}$ and $\ln |y_{u,j}/y_{l,j}| = \ln |z_{ij}/(x_{l,i} y_{l,j})|$. So the derivative writes

$$\begin{aligned} f_{z,ij,d}(z) = \quad & p_{0,x_i} p_{0,y_j} (\ln |y_{u,j}/y_{l,j}| + z_{ij} z_{ij}^{-1} - x_{l,i} x_{l,i}^{-1}) \\ & + p_{0,x_i} r_{y_j} ((y_{u,j} - y_{l,j}) + z_{ij} x_{l,i}^{-1} - \tfrac{1}{2} x_{l,i} 2 y_{u,j} x_{l,i}^{-1}) \\ & + \tfrac{1}{2} r_{x_i} p_{0,y_j} (-2 z_{ij} (y_{u,j}^{-1} - y_{l,j}^{-1}) + z_{ij}^2 x_{l,i} z_{ij}^{-2} - x_{l,i}^2 x_{l,i}^{-1}) \\ & + \tfrac{1}{2} r_{x_i} r_{y_j} (2 z_{ij} \ln |y_{u,j}/y_{l,j}| + z_{ij}^2 z_{ij}^{-1} - \tfrac{1}{2} x_{l,i}^2 2 y_{u,j} x_{l,i}^{-1}). \end{aligned} \tag{A.33}$$

and can be rewritten as

$$\begin{aligned} f_{z,ij,d}(z) = \quad & p_{0,x_i} p_{0,y_j} \ln |y_{u,j}/y_{l,j}| \\ & + p_{0,x_i} r_{y_j} (y_{u,j} - y_{l,j}) \\ & - r_{x_i} p_{0,y_j} z_{ij} (y_{u,j}^{-1} - y_{l,j}^{-1}) \\ & + r_{x_i} r_{y_j} z_{ij} \ln |y_{u,j}/y_{l,j}|, \end{aligned} \tag{A.34}$$

where $z_{ij}(y_{u,j}^{-1} - y_{l,j}^{-1})$ can be replaced by $-(x_{u,i} - x_{l,i})$.

Division

Let $Z = X/Y$. For division integration of probability for joint bins has to be performed separately for each quadrant, as can be seen in Figure 2.3. In this section, integration for quadrant 1 ($z \in \langle 0, \infty \rangle$) is derived. The integration boundaries for joint bin (i, j) are defined as

$$\begin{aligned} y_{u,j} &= \max(y_j, \min(y_{j+1}, x_{i+1}/z)) \\ x_{u,i} &= \max(x_i, \min(x_{i+1}, z y_{u,j})) \\ y_{l,j} &= \max(y_j, \min(y_{j+1}, x_{l,i}/z)) \\ x_{l,i} &= \max(x_i, \min(x_{i+1}, z y_j)) \\ z_{ij} &= x_{u,i}/y_{u,j} = x_{l,i}/y_{l,j}. \end{aligned} \tag{A.35}$$

Equation (A.5) for sub-area d can be written as

$$F_{z,ij,d}(z) = \int_{y=y_{l,j}}^{y_{u,j}} \int_{x=x_{l,i}}^{z_{ij} y} f_{x,i}(x) f_{y,j}(y) \, \mathrm{d}x \, \mathrm{d}y. \tag{A.36}$$

Integration with respect to x yields

$$F_{z,ij,d}(z) = \int_{y=y_{l,j}}^{y_{u,j}} \left[p_{0,x_i}(z_{ij}y - x_{l,i}) + \tfrac{1}{2}r_{x_i}\left((z_{ij}y)^2 - x_{l,i}^2\right)\right] f_{y,j}(y)\,\mathrm{d}y. \tag{A.37}$$

Integration with respect to y yields

$$\begin{aligned} F_{z,ij,d}(z) = \big[\, & p_{0,x_i}p_{0,y_j}(\tfrac{1}{2}z_{ij}y^2 - x_{l,i}y) + p_{0,x_i}r_{y_j}(\tfrac{1}{3}z_{ij}y^3 - \tfrac{1}{2}x_{l,i}y^2) \\ & + \tfrac{1}{2}r_{x_i}p_{0,y_j}(\tfrac{1}{3}z_{ij}^2y^3 - x_{l,i}^2y) + \tfrac{1}{2}r_{x_i}r_{y_j}(\tfrac{1}{4}z_{ij}^2y^4 - \tfrac{1}{2}x_{l,i}^2y^2)\big]_{y=y_{l,j}}^{y_{u,j}}. \end{aligned} \tag{A.38}$$

Inserting integration boundaries yields

$$\begin{aligned} F_{z,ij,d}(z) = \; & p_{0,x_i}p_{0,y_j}(\tfrac{1}{2}z_{ij}(y_{u,j}^2 - y_{l,j}^2) - x_{l,i}(y_{u,j} - y_{l,j})) \\ & + p_{0,x_i}r_{y_j}(\tfrac{1}{3}z_{ij}(y_{u,j}^3 - y_{l,j}^3) - \tfrac{1}{2}x_{l,i}(y_{u,j}^2 - y_{l,j}^2)) \\ & + \tfrac{1}{2}r_{x_i}p_{0,y_j}(\tfrac{1}{3}z_{ij}^2(y_{u,j}^3 - y_{l,j}^3) - x_{l,i}^2(y_{u,j} - y_{l,j})) \\ & + \tfrac{1}{2}r_{x_i}r_{y_j}(\tfrac{1}{4}z_{ij}^2(y_{u,j}^4 - y_{l,j}^4) - \tfrac{1}{2}x_{l,i}^2(y_{u,j}^2 - y_{l,j}^2)). \end{aligned} \tag{A.39}$$

The first derivative of Equation (A.39) with respect to z_{ij} is its corresponding PDF. The variables dependent on z_{ij} are $x_{u,i} = z_{ij}y_{u,j}$ and $y_{l,j} = x_{l,i}/z_{ij}$. So the derivative writes

$$\begin{aligned} f_{z,ij,d}(z) = \; & p_{0,x_i}p_{0,y_j}(\tfrac{1}{2}y_{u,j}^2 + \tfrac{1}{2}x_{l,i}^2z_{ij}^{-2} - x_{l,i}^2z_{ij}^{-2}) \\ & + p_{0,x_i}r_{y_j}(\tfrac{1}{3}y_{u,j}^3 + \tfrac{1}{3}2x_{l,i}^3z_{ij}^{-3} - \tfrac{1}{2}2x_{l,i}^3z_{ij}^{-3}) \\ & + \tfrac{1}{2}r_{x_i}p_{0,y_j}(\tfrac{1}{3}2z_{ij}y_{u,j}^3 + \tfrac{1}{3}x_{l,i}^3z_{ij}^{-2} - x_{l,i}^3z_{ij}^{-2}) \\ & + \tfrac{1}{2}r_{x_i}r_{y_j}(\tfrac{1}{4}2z_{ij}y_{u,j}^4 + \tfrac{1}{4}2x_{l,i}^4z_{ij}^{-3} - \tfrac{1}{2}2x_{l,i}^4z_{ij}^{-3}). \end{aligned} \tag{A.40}$$

and can be rewritten as

$$\begin{aligned} f_{z,ij,d}(z) = \; & \tfrac{1}{2}p_{0,x_i}p_{0,y_j}(y_{u,j}^2 - y_{l,j}^2) \\ & + \tfrac{1}{3}p_{0,x_i}r_{y_j}(y_{u,j}^3 - y_{l,j}^3) \\ & + \tfrac{1}{3}r_{x_i}p_{0,y_j}z_{ij}(y_{u,j}^3 - y_{l,j}^3) \\ & + \tfrac{1}{4}r_{x_i}r_{y_j}z_{ij}(y_{u,j}^4 - y_{l,j}^4). \end{aligned} \tag{A.41}$$

A.2 Performance examples

As a showcase, two examples from the literature are chosen to reproduce the results using piecewise linear PDFs, and to compare the performance of the different calculation methods.

If an analytical solution of a stochastic problem is not available, or hard to achieve, then Monte Carlo simulation is a frequently used method to get an answer to the problem. Calculations with piecewise linear PDFs can be a good alternative for Monte Carlo simulations. From the literature, two examples are selected to demonstrate the performance differences between these two methods. In both selected papers, a Monte Carlo simulation is performed as a reference for an approximate analytical solution.

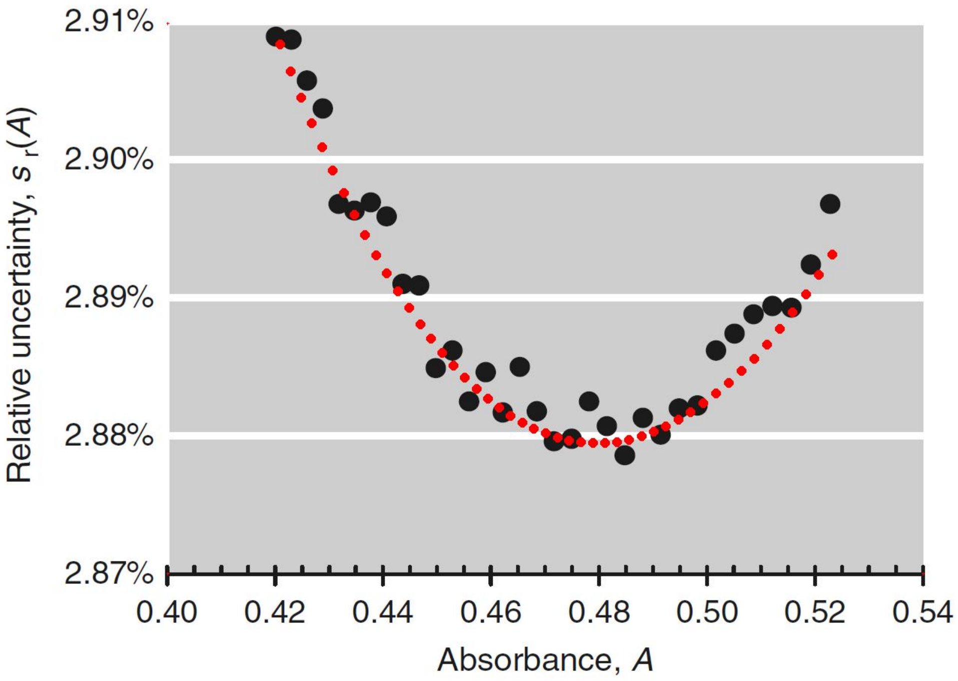

Figure A.2: *Copied from Meija [2010] (black), showing the results of evaluating equation A.42 by using Monte Carlo simulation. The red dots are the results of the piecewise linear* PDF *calculations.*

A.2.1 Example 1: lowest relative uncertainty

Meija [2010] used two methods, Monte Carlo simulation (MC) and analytical derivation, to find the lowest relative uncertainty (i.e., coefficient of variation) of a formula calculating the light absorbance value. We compared the performance of these methods with our PDF calculations. The formula Meija used is

$$A = -\log_{10}\frac{I}{I_0}, \tag{A.42}$$

where A is the absorbance value, and I and I_0 are light intensities with a standard deviation of 0.01. To find the optimal ratio of I and I_0 the mean value of I is varied between 0.3 and 0.4 and I_0 has a fixed mean of 1. These variables are assumed to be Gaussian distributed.

The aim is to find the lowest value of relative uncertainty σ_A/A where σ_A is the standard deviation of A. Figure A.2 shows the results (black dots) of the evaluation of equation A.42, as calculated by *Meija* [2010]. The red dots are the results of the piecewise linear PDF calculations. The lowest point of the graph is at $A \approx 0.48$, which is approximately the same value as found by *Meija* [2010]: 0.48 with the Monte Carlo simulation, and 0.482 with the analytical solution.

It should be noted that the relative uncertainty of A is sensitive to the discretization of the PDFs of I and I_0. They could be described sufficiently accurate by at least

40 bins. The truncation of the PDFs was performed two sided at five times the standard deviation. All bins were of equal width.

A.2.2 Example 2: uncertainty of calculated ratios

Holmes and Buhr [2007] addressed the uncertainty of quantities calculated from laboratory measures. They developed an analytical equation to calculate the coefficient of variation (CV) of the ratio of two Gaussian distributed RVs. This equation is defined as:

$$CV_R \cong \frac{\sqrt{CV_X^2 + CV_Y^2 + 3CV_Y^2 CV_X^2 + 8CV_Y^4}}{1 + CV_Y^2}. \tag{A.43}$$

The ratio of the RVs is written as: $R = X/Y$, where X and Y are known independent and Gaussian distributed RVs. The CV of an RV is defined as the quotient of the standard deviation and the average value of this RV. Thus $CV_X = \sigma_X/\mu_X$ and $CV_Y = \sigma_Y/\mu_Y$, where σ and μ are the standard deviation and average value of their respective variables. In this example, the average values of X and Y are 650 and 0.14, respectively. The variances of these values are derived from the given CVs to be tested. These values are $0.05, 0.10, ..., 0.30$ for X and $0.01, 0.02, ..., 0.30$ for Y.

In figure A.3, the CVs of R are plotted against the CVs of Y for fixed CV values of X. As can be seen, the values of the PDF calculations (red) correspond quite well with the results of the Monte Carlo simulation (MC). The values of the MC are used by *Holmes and Buhr* [2007] as a benchmark for equation A.43.

The PDFs were discretized in 50 bins, and, like *Holmes and Buhr* [2007] did, two sided truncated at three times the standard deviation. To reproduce the results of Holmes and Buhr, it was important to pursue this truncation in the same way.

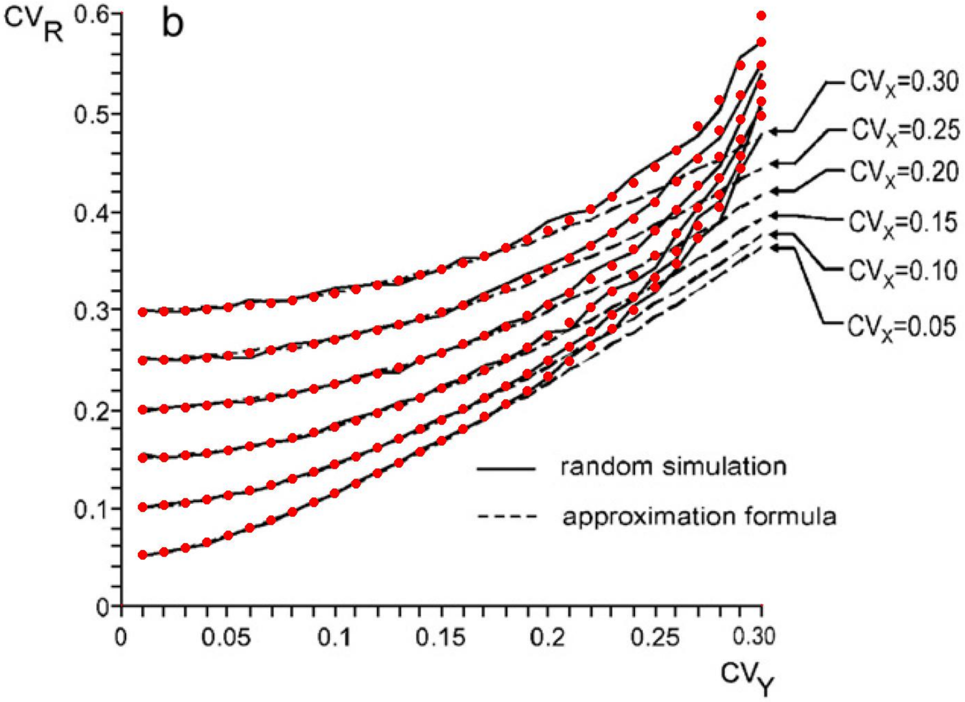

Figure A.3: *Copied from Holmes and Buhr [2007] (black). The solid black lines are the result of a Monte Carlo simulation of Holmes and Buhr [2007], the dashed lines are the results of their improved formula, and the red dots are the results of our* PDF *calculations.*

B

Mode of conditional joint distributions

B.1 Mode of the joint distribution of elementary operations

Finding the mode of a multiple dimensional joint probability density function (PDF) of independent random variables (RVs) is straight forward. The mode is found at the position formed by the modes of the marginal distributions. When the joint PDF is subject to any constraint, finding the mode is less obvious. This section contains the derivations of finding the mode when the result of an elementary operation of the marginal PDFs is conditional on a fixed value. All RVs are described by piecewise linear PDFs.

Hereafter, X and Y are known independent RVs and Z is the resulting RV of an elementary operation $(+ - \times /)$. For every value $x \in X$ and $y \in Y$ the probability density of the joint distribution can be calculated as

$$p(x, y) = f_{x_i}(x) f_{y_j}(y), \tag{B.1}$$

where $f_x(x)$ and $f_y(y)$ are the PDFs of X and Y, respectively. The subscripts i and j denote the bin numbers of the piecewise linear PDFs. The PDFs are defined as

$$f_{x_i}(x) = p_{0,x_i} + r_{x_i} x \tag{B.2}$$

$$f_{y_j}(y) = p_{0,y_j} + r_{y_j} y, \tag{B.3}$$

where p_{0,x_i} and p_{0,y_j} are the probability densities at $x = 0$ for bin i and $y = 0$ for bin j, respectively, and r_{x_i} and r_{y_j} are constant values.

Applying elementary operations, x can be written as a function of z and y as

$$x = g(y, z). \tag{B.4}$$

Inserting Eqs. (B.2)–(B.4) into Eq. (B.1) yields

$$p(y, z) = (p_{0,x_i} + r_{x_i} g(y, z))(p_{0,y_j} + r_{y_j} y). \tag{B.5}$$

The extreme values of $p(g(y, z), y)$ for a certain value of z can be found by taking the first derivative with respect to y, which writes

$$\frac{\mathrm{d}p(y, z)}{\mathrm{d}y} = p_{0,x_i} r_{y_j} + r_{x_i} p_{0,y_j} \frac{\mathrm{d}g(y, z)}{\mathrm{d}y} + r_{x_i} r_{y_j} \frac{\mathrm{d}g(y, z)y}{\mathrm{d}y}. \tag{B.6}$$

Setting this function equal to 0 and solve it for y yields the coordinates (x, y) with an extreme value for $p(x, y)$. Since this function only holds within the domain of the joint bin (i, j), the value of y must satisfy the constraint $y \in [y_j, y_{j+1}]$, where y_j and y_{j+1} are the boundaries of the bin j of Y. Equivalently, x is constrained to $x \in [x_i, x_{i+1}]$. All bins which are intersected by the line $x = g(y, z)$ have to be evaluated to find the mode.

In the next sections this method is applied to four elementary operations.

B.1.1 Summation

Let $Z = X+Y$, thus $g(y,z) = z-y$. The first derivative with respect to y of Eq. (B.6) yields:

$$\begin{aligned}\frac{\mathrm{d}p(y,z)}{\mathrm{d}y} =& p_{0,x_i}r_{y_j} + r_{x_i}p_{0,y_j}\frac{\mathrm{d}(z-y)}{\mathrm{d}y} \\ =& p_{0,x_i}r_{y_j} - r_{x_i}p_{0,y_j} + r_{x_i}r_{y_j}z - 2r_{x_i}r_{y_j}y.\end{aligned} \tag{B.7}$$

Setting this function to 0 and solve it for y yields

$$y = (p_{0,x_i}r_{y_j} - r_{x_i}p_{0,y_j} + r_{x_i}r_{y_j}z)/(2r_{x_i}r_{y_j}). \tag{B.8}$$

B.1.2 Subtraction

Let $Z = X-Y$, thus $g(y,z) = z+y$. The first derivative with respect to y of Eq. (B.6) yields

$$\begin{aligned}\frac{\mathrm{d}p(y,z)}{\mathrm{d}y} =& p_{0,x_i}r_{y_j} + r_{x_i}p_{0,y_j}\frac{\mathrm{d}(z+y)}{\mathrm{d}y} + r_{x_i}r_{y_j}\frac{\mathrm{d}(z+y)y}{\mathrm{d}y} \\ =& p_{0,x_i}r_{y_j} + r_{x_i}p_{0,y_j} + r_{x_i}r_{y_j}z + 2r_{x_i}r_{y_j}y.\end{aligned} \tag{B.9}$$

Setting this function to 0 and solve it for y yields

$$y = (p_{0,x_i}r_{y_j} + r_{x_i}p_{0,y_j} + r_{x_i}r_{y_j}z)/(-2r_{x_i}r_{y_j}). \tag{B.10}$$

B.1.3 Multiplication

Let $Z = XY$, thus $g(y,z) = z/y$. The first derivative with respect to y of Eq. (B.6) yields

$$\begin{aligned}\frac{\mathrm{d}p(y,z)}{\mathrm{d}y} =& p_{0,x_i}r_{y_j} + r_{x_i}p_{0,y_j}\frac{\mathrm{d}(z/y)}{\mathrm{d}y} + r_{x_i}r_{y_j}\frac{\mathrm{d}(z/y)y}{\mathrm{d}y} \\ =& p_{0,x_i}r_{y_j} - r_{x_i}p_{0,y_j}zy^{-2}.\end{aligned} \tag{B.11}$$

Setting this function to 0 and solve it for y yields

$$y = \pm\sqrt{\frac{r_{x_i}p_{0,y_j}z}{p_{0,x_i}r_{y_j}}}. \tag{B.12}$$

B.1.4 Division

Let $Z = X/Y$, thus $g(y,z) = zy$. The first derivative with respect to y of Eq. (B.6) yields

$$\begin{aligned}\frac{\mathrm{d}p(y,z)}{\mathrm{d}y} =& p_{0,x_i}r_{y_j} + r_{x_i}p_{0,y_j}\frac{\mathrm{d}(zy)}{\mathrm{d}y} + r_{x_i}r_{y_j}\frac{\mathrm{d}(zy)y}{\mathrm{d}y} \\ =& p_{0,x_i}r_{y_j} + r_{x_i}p_{0,y_j}z + 2r_{x_i}r_{y_j}zy.\end{aligned} \tag{B.13}$$

Setting this function to 0 and solve it for y yields

$$y = (p_{0,x_i}r_{y_j} + r_{x_i}p_{0,y_j}z)/(-2r_{x_i}r_{y_j}z). \tag{B.14}$$

C

Theories connected

In the literature, inference in probability theory is described using different styles of notation and terminology. The derivation of rules or axioms for processing probability data are found in the theory of the Boolean algebra, the set theory, and the conventional probability theory. The Bayesian belief theory, as can be seen as founded by Dempster and Shafer [*Dempster*, 1966, 1967; *Shafer*, 1976], is important in the literature and is used in describing the inference in Bayesian networks (BNs) or directed acyclic graphs (DAGs). All these branches of theory do have their own notation and rules, but do often describe the same phenomena. Hereafter, a short overview is given of the connection between the different styles as used in different fields of the probability literature. This is not an exhaustive description of the different theories, but only a description of the common parts to be able to see the connections.

C.1 The connection

Let the variables A and B describe two propositions (statements) or events, and let S define all possible events or the event space. If the variables are Boolean they take only two values: 0 or 1, or, respectively, false or true. In the set theory they can take any value. To the occurrence of an event a probability can be assigned, which makes a connection between Boolean or set theory, and conventional probability theory. This probability is always greater then or equal to 0, also known as the first Kolmogorov axiom [*Kjærulff and Madsen*, 2012, p. 40]. The notation $\mathrm{Pr}(A)$ is defined as the probability of event A taking place.

In the Boolean algebra [e.g. *Jaynes*, 2003, p. 9 ff] three basic operations on propositions are defined. The first operation is the logical product, logical AND, or conjunction, which is symbolically written as AB, $A \cdot B$, $A \wedge B$, or just A AND B. This means that if both A and B are true, then AB is true, otherwise AB is false. The second operation is the logical sum, logical OR, or disjunction, which is denoted by $A + B$, $A \vee B$, or A OR B. This disjunction is defined as true if either A or B or both are true, thus only if both A and B are false then the disjunction evaluates to false. The last operation is the logical NOT, or negation written as $\overline{A}$ or $\neg A$. This means that if A is true, then $\overline{A}$ is false, and the reverse. These three operations are sufficient to perform all possible logical operations [*Jaynes*, 2003, p. 15]. Moreover, through the duality property $\overline{AB} = \overline{A} + \overline{B}$ it is even possible to define all operations with only the negation operator and one of the conjunction or disjunction operators.

In the set theory [e.g. *Papoulis and Pillai*, 2002, p. 15 ff; *Kjærulff and Madsen*, 2012, p. 40], objects or elements and operations are defined which are comparable to those in the Boolean algebra but using different notations. In this context, set S is the event space or sample space, and A and B are subsets of S. The event space S contains elements which may be discrete or continuous. A frequently used tool is a Venn diagram, which is a useful tool for visually reasoning. Figure C.1 shows an example with event space S and two subsets A and B.

The area within a subset can contain discrete elements or denotes a continuous

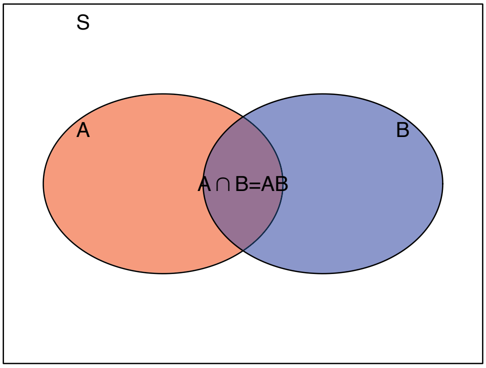

Figure C.1: *Example of a Venn diagram with subset A and B in event space S.*

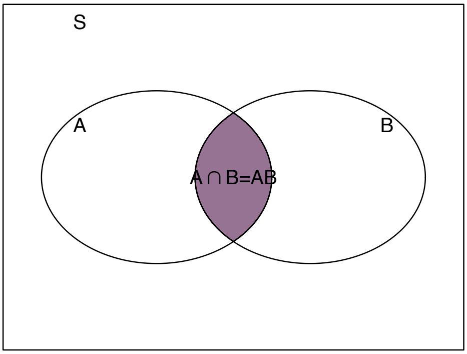

Figure C.2: *Venn diagram showing the intersection of subset A and B in event space S.*

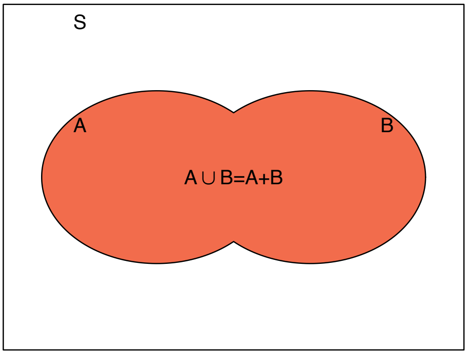

Figure C.3: *Venn diagram showing the union of subset A and B in event space S.*

area. In case of a random process, to every discrete element or every sub-area a probability is assigned. In case of a continuous space S, every point in space can be assigned a probability density. A Venn diagram may not be suitable to describe all elements of probability theory [*Jaynes*, 2003, p. 48], but it suits all the needs for the problems at hand in Chapter 5. For set theory, the same three basic operations as for Boolean algebra are defined. Firstly, the intersection operation AB or $A \cap B$, related to the conjunction operation, describes the intersection of set A and B. So the set AB contains all elements which are in set A and in set B. This is in Figure C.2 denoted by the purple area. Secondly, the union operation $A+B$ or $A \cup B$, related to the disjunction operator, denotes the union of the sets A and B. So every element in set A, in set B or in both sets is an element of set $A+B$. In Figure C.3 this is the red area. An important property of sets is that all elements within a set are mutually exclusive. In case of events, this means that two mutually exclusive events can not happen at the same time. So even when A and B are not mutually exclusive, which means that the intersection AB is not the empty set $\{\emptyset\}$, still the set $A+B$ contains only distinct elements. Thirdly, related to the negation, the complement of a set A is written as $\overline{A}$, A^c, or A'. So set $\overline{A}$ contains all elements of S excluded the elements

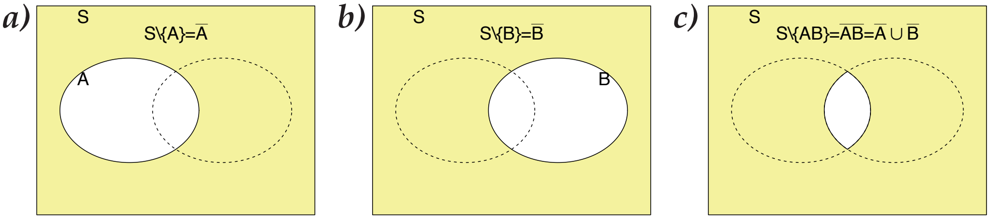

Figure C.4: *Venn diagrams showing the duality property of sets. Figure a) depicts the complement of A, figure b) the complement of B, and figure c) shows that the union of the complements is equal tot the complement of the intersection of subset A and B.*

of A, also written as $S \setminus \{A\}$ (Figure C.4a). Equivalent to the Boolean algebra, in the set theory the duality property or De Morgan's law [*Papoulis and Pillai*, 2002, p. 18] is defined as $\overline{AB} = \overline{A} \cup \overline{B}$. In Figure C.4 this is depicted as $\overline{A}$ (C.4a), $\overline{B}$ (C.4b), and the union $\overline{A} \cup \overline{B}$ (C.4c).

Further, in conventional probability theory the concepts of the former theories are applied, or rather, probabilities are assigned to the propositions or events of these theories. In the Boolean algebra, if A is a proposition, or Boolean expression, which can be either true or false with a certain probability, then $\Pr(A)$ is the probability of A being true. Also, if in the set theory a probability is assigned to the elements (or areas) of event space S, then $\Pr(A)$ is the probability of an element s of S $(s \in S)$ being in set A, or $\Pr(A) = \sum_{s \in A} \Pr(s)$. So $\Pr(A)$ is equal to the sum of the probabilities of all elements s which are an element of A. Although S and A can contain much more elements than two, the probability can still be written as a Boolean expression. Therefore, we can define the proposition $'s$ *is an element of* A' or $s \in A$, which is either true or false with a certain probability. So we can write $\Pr(A) = \Pr('s$ *is an element of* $A')$ in which the left hand side is written as the probability of a set, and the right hand side as the probability of a Boolean expression. With probabilities assigned to all elements of set S, this set can be seen as a random variable (RV), or, in other words, the RV is a function which assigns probabilities to the elements of set S [*Papoulis and Pillai*, 2002, p. 15]. An important axiom of the probability theory is that the probability of all elements s of S sum to 1, also known as the second Kolmogorov axiom [*Kjærulff and Madsen*, 2012, p. 40].

When defined in this way, S seems to be a single RV. It is nevertheless easy to define S as the combination of multiple variables. Let X and Y be random variables with elements x_i with $i = 1 \dots m$, and y_j with $j = 1 \dots n$, respectively, and let all elements of S be defined as $s_{ij} = (x_i, y_j)$. Now, the joint probability [*Bishop*, 2006, p. 13; *Papoulis and Pillai*, 2002, p. 169] of event $X = x_i$ and $Y = y_j$ is defined as the probability of both events happening simultaneously, written as $\Pr(X = x_i, Y = y_j)$. This means the probability of RV X taking the value x_i and RV Y taking the value y_j simultaneously. If we define set $A_i = \{s_{i1}, \dots, s_{in}\}$ and set $B_j = \{s_{1j}, \dots, s_{mj}\}$, then the intersection $A_i \cap B_j = \{s_{i,j}\} = \{(x_i, y_j)\}$, hence $\Pr(A_i \cap B_j) = \Pr(x_i, y_j)$. So, the definition of the joint probability is clearly related to intersection of the set theory and the conjunction of the Boolean algebra.

Therefore, the notations of the set theory and the Boolean algebra are found in the probability literature as well.

In consistence with the union operation, *the sum rule of probability* [*Jaynes*, 2003, p. 30 ff], or *rule of total probability* [*Kjærulff and Madsen*, 2012, p. 44], also known as the third Kolmogorov axiom [*Kjærulff and Madsen*, 2012, p. 40], is defined. When A and B are mutual exclusive events, then the probability of $A + B$ equals to the sum of the probabilities of A and B, $\Pr(A) + \Pr(B)$. When A and B are not mutual exclusive, then the probability writes

$$\Pr(A + B) = \Pr(A) + \Pr(B) - \Pr(AB). \tag{C.1}$$

A summary of equivalent notations of the three operations is, for the conjunction or union operation

$$\begin{aligned}\Pr(A + B) &= \Pr(A \text{ OR } B) = \Pr(A \vee B) = \Pr(A \cup B)\\ &= \Pr(A) + \Pr(B) - \Pr(AB),\end{aligned} \tag{C.2}$$

for the disjunction or intersection operation or joint distribution

$$\Pr(AB) = \Pr(A \cdot B) = \Pr(A \text{ AND } B) = \Pr(A \wedge B) = \Pr(A \cap B) = \Pr(A, B), \tag{C.3}$$

and for the negation or complement

$$\Pr(\overline{A}) = \Pr(\neg A) = \Pr(A^c) = \Pr(A') = 1 - \Pr(A). \tag{C.4}$$

C.2 Conditional probability and Bayes's rule

By application of the theory of the former section, the interpretation of conditional probability can easily be explained and understood. The conditional probability [*Kjærulff and Madsen*, 2012, p. 41] is a very important concept in probability theory, and especially in Bayesian inference. The conditional probability is written as $\Pr(A|B)$, which means the probability of event A given that event B has occurred. Or in terms of sets, given that s is an element of B, what is the probability that s is an element of A too. In Figure C.5 this concept is graphically displayed for $\Pr(A|B)$ and $\Pr(B|A)$. In these figures, the probability of an element s being in the shaded area is 1, because this is given. So the probability of all elements s within this area have to be divided by the total probability of the shaded area, which obviously is 1. From Figure C.5 it can be seen that we can write

$$\Pr(A|B) = \frac{\Pr(AB)}{\Pr(B)} \tag{C.5}$$

$$\Pr(B|A) = \frac{\Pr(AB)}{\Pr(A)} \tag{C.6}$$

Combining these equations yields

$$\Pr(AB) = \Pr(A|B)\Pr(B) = \Pr(B|A)\Pr(A), \tag{C.7}$$

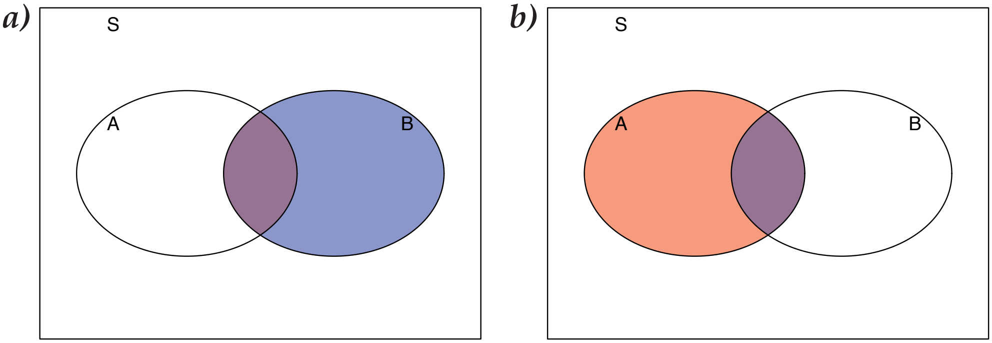

Figure C.5: *Graphical interpretation of the conditional probability. The interpretation of the intersection of A and B is in figure a):* $\Pr(A|B)$*, and in figure b):* $\Pr(B|A)$*.*

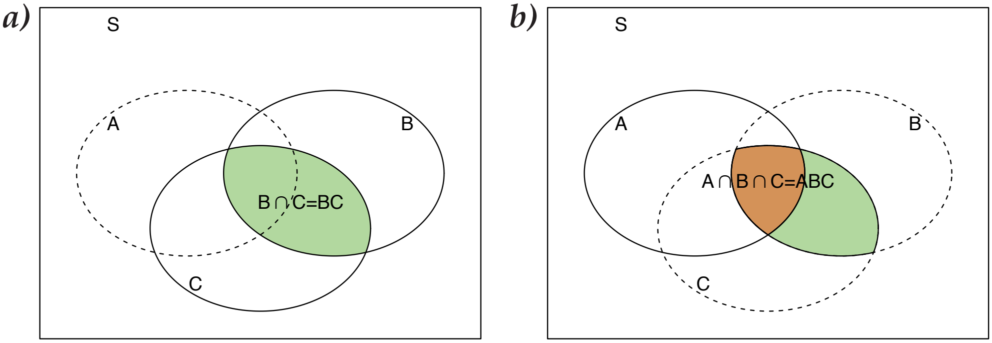

Figure C.6: *Graphical interpretation of the chain rule. Figure a) can be interpreted as* $\Pr(BC) = \Pr(C|B)\Pr(B) = \Pr(B|C)\Pr(C)$*. Figure b) depicts* $\Pr(A|BC)\Pr(BC)$*, of which* $\Pr(BC)$ *can be further expanded as in figure a).*

which is called *the fundamental rule of probability* [*Bishop*, 2006, p. 13 ff; *Kjærulff and Madsen*, 2012, p. 42,54], or the *factorization* of the joint distribution of AB [*Kjærulff and Madsen*, 2012, p. 48]. Rewriting the fundamental rule of Equation (C.7) yield Bayes' Theorem [*Kjærulff and Madsen*, 2012, p. 54]

$$\Pr(A|B) = \frac{\Pr(AB)}{\Pr(B)} = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}, \tag{C.8}$$

and equivalently for $\Pr(B|A)$. Recursive application of the fundamental rule yields *the chain rule* [*Kjærulff and Madsen*, 2012, p. 62]. So the factorization of the joint distribution $\Pr(ABC)$ writes

$$\Pr(ABC) = \Pr(A|BC)\Pr(BC) = \Pr(A|BC)\Pr(B|C)\Pr(C), \tag{C.9}$$

in which A, B, and C may be freely interchanged to arrive at different expressions, but all describing the same joint distribution. In Figure C.6 the chain rule is graphically shown with a Venn-diagram. In the set theory the result of an operation on two sets always yields one new set. But the reverse can also be stated, that one set

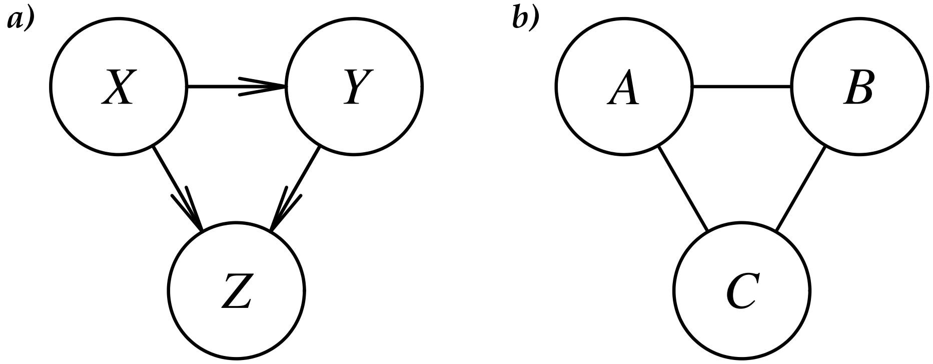

Figure C.7: Simple Bayesian graph (left) and Markov network (right) with three stochastic nodes.

always can be written as the result of an operation on two sets. With this mind, Eq. C.7 and C.9 can easily be seen as equivalent.

If event A is independent of B then the conditional probability of A given B can be written as $\mathrm{Pr}(A|B) = \mathrm{Pr}(A)$. This is not equal to A and B being mutual exclusive.

C.3 Probability potentials

In the literature, inference in Bayesian networks/graphs is often described in terms of potentials, instead of probability density functions. The notation and terminology diverges strongly from the common practice in other areas of probability literature. The connection between these areas is described here.

A Bayesian graph (Figure C.7a), or directed acyclic graph (DAG), is a graphical representation of a joint probability distribution containing conditional distributions. In a Bayesian graph, each node represents a random variable, and the nodes are connected by arrows (or edges). These arrows between the nodes represent the conditional dependencies between the variables. Together, the total graph describes a joint distribution where the probability, or probability density, depends on the values of each node. For each variable, the probability distribution of this variable depends on the values of its parents, and can be described by conditional distributions. In Figure C.7, a Bayesian graph with nodes X, Y and Z is shown. There, the set of parents of Y is $\mathrm{pa}(Y) = \{X\}$, where $\mathrm{pa}(Y)$ denotes the parents of Y, and the set of parents of Z is $\mathrm{pa}(Z) = \{X, Y\}$.

In contrast to the Bayesian graph an undirected graph or Markov network can be used to describe a probability distribution. In Figure C.7b such a graph is depicted. The difference between the two networks is the dependence structure of variables, the edges in the Markov network do not contain arcs. This implies that each node has no parents and is not defined as a conditional distribution.

A set of specific values of the domain of the variables in a probabilistic graph is called a state. The collection of all states of a graph is often denoted by Ω, and the collection of states of one variable, say X, is defined as Ω_X. If the random variable

X is interpreted as a function then Ω_X is the domain of this function. If the state of a variable is observed this state is often called evidence. If a node in a DAG is observed then the dependency structure may change. This is known as directional or d-separation [e.g. *Pearl*, 1993; *Kjærulff and Madsen*, 2012, p. 33].

In both the directed and undirected graphs the probability functions of the nodes are denoted by probability potential. A (probability) potential is defined as a non-negative function but is not necessarily a probability distribution, since it not always integrates to 1 [*Kjærulff and Madsen*, 2012, p. 46]. Nevertheless, a potential can be turned into a probability distribution by normalization. This is often the case in an undirected graphical model, or Markov network (Figure C.7b), but in a DAG the nodes usually represent conditional distributions which are normalized [*Bishop*, 2006, p. 386]. For instance, the potential of node C in Figure C.7b, given the state of node A and B can be written as $\mathrm{Pr}(C, A = a, B = b)$. Unless A and B are degenerate random variables (with 0-variances), the potential of C is not a probability distribution but is a subset of the joint distribution of the total network. Hence the function does not integrate to 1. For the graph in Figure C.7a, the joint distribution may be written as $\mathrm{Pr}(Z, X, Y) = \mathrm{Pr}(Z|X, Y)\,\mathrm{Pr}(Y|X)\,\mathrm{Pr}(X)$, where the factorization is explicitly given by the direction of the arrows. Each factor on the right hand side of this equation coincides with the probability distribution of one node. So given the state of X and Y, the potential of Z writes $\mathrm{Pr}(Z|X = x, Y = y)$. Due to normalization, this potential is a probability distribution. It is common practice to denote a potential by a Greek letter, so $\zeta = \mathrm{Pr}(Z|X = x, Y = y)$.

The application of potentials in the description of Bayesian networks has its own operations and notations. This yields a very compact notation which may need some explanation. The domain of a potential is defined as the set of all variables involved [*Cabañas et al.*, 2014, p. 99], which is the union of the variable and its parents. So in the example of Figure C.7a, the domain of ζ is $\mathrm{dom}(\zeta) = \{Z, \mathrm{pa}(Z)\} = \{X, Y, Z\}$, the domain of ψ, which is the potential of Y, is $\mathrm{dom}(\psi) = \{X, Y\}$, and the domain of ξ, the potential of X, is $\mathrm{dom}(\xi) = \{X\}$. Combination of potentials is written as $\xi \otimes \psi$, where $\otimes$ denotes a point-wise multiplication [*Cinicioglu and Shenoy*, 2009]. The domain of such a potential is the union of the domains of the individual potentials, so $\mathrm{dom}(\xi \otimes \psi) = \mathrm{dom}(\xi) \cup \mathrm{dom}(\psi) = \{X, Y\}$. For a specific state (x, y), the potential is written as $(\xi \otimes \psi)(x, y)$, which is equivalent to $\mathrm{Pr}(x, y) = \mathrm{Pr}(y|x)\,\mathrm{Pr}(x)$ in the above example.

Marginalization is an operation on a conditional distribution where a marginal variable is integrated out of the distribution (see Section C.4). Marginalization of potentials is denoted by $\psi' = (\xi \otimes \psi)^{-X}$, where ψ' is a new potential or function of Y, in this example with X marginalized out. An equivalent expression is $\mathrm{Pr}(y) = \sum_{x \in X} \mathrm{Pr}(y|x)\,\mathrm{Pr}(x)$ for discrete X, or $\mathrm{Pr}(y) = \int_x p(y|x)p(x)\mathrm{d}x$ for a continuous variable X.

Another frequently used operation is the projection [*Cinicioglu and Shenoy*, 2009], which can be seen as the complement of a marginalization. The projection is an operation on the state of a potential. Let $\mathbf{v}$ be a state of the total graph, and let Ω_X, Ω_Y and Ω_Z be the domain of X, Y and Z, respectively. Now the potentials of X, Y and Z with state $\mathbf{v}$ can be written as $\xi(\mathbf{v}^{\downarrow\Omega_X})$, $\psi(\mathbf{v}^{\downarrow\Omega_Y})$ and $\zeta(\mathbf{v}^{\downarrow\Omega_Z})$, respectively. So the projection can be seen as a selection of elements of the state $\mathbf{v}$ needed for a specific potential. Applied to a combination of potentials this yields $(\xi \otimes \psi \otimes \zeta)(\mathbf{v}) = \xi(\mathbf{v}^{\downarrow\Omega_X})\psi(\mathbf{v}^{\downarrow\Omega_Y})\zeta(\mathbf{v}^{\downarrow\Omega_Z})$.

C.4 Marginalization, elimination of nuisance parameters

Making inference in a Bayesian network (BN) involves updating of the marginal distributions, given observations of some variables. Such a network may contain a large number of marginal distributions, but one may be interested in updating only a few of these distributions. The variables which are, for the moment, of no interest are the so called nuisance parameters. When multiple marginal parameters are of interest, then updating all marginal distributions at the same time may be intractable. In such a case, a number of marginal distributions can be treated as nuisance parameters, retaining only a small number of distributions to be updated. The nuisance parameters can be marginalized or integrated out, which decreases the size and the complexity of the network [*Kjærulff and Madsen*, 2012, p. 115; *Held and Bové*, 2013, p. 200; *Gelman et al.*, 2014, p. 63]. This marginalization can be seen as combining the information of the nuisance parameters and transfer this information to the variables of interest.

A Bayesian network can be written as a joint distribution and a factorization of this distribution (see Section C.2). Let $p(\theta_1, \theta_2, y)$ be a joint distribution of three arbitrary variables θ_1, θ_2 and y. Any joint distribution of three variables can be factorized in nine different ways. Three factorizations of $p(\cdot)$ are

$$\begin{aligned} p(\theta_1, \theta_2, y) &= p(\theta_1|\theta_2, y)p(\theta_2, y) \\ &= p(\theta_1, \theta_2|y)p(y), \end{aligned} \tag{C.10}$$

and by applying the chain rule $p(\theta_2, y) = p(\theta_2|y)p(y) = p(y|\theta_2)p(\theta_2)$ this yields

$$\begin{aligned} p(\theta_1, \theta_2, y) &= p(\theta_1|\theta_2, y)p(\theta_2|y)p(y) \\ &= p(\theta_1|\theta_2, y)p(y|\theta_2)p(\theta_2) \\ &= p(\theta_1, \theta_2|y)p(y). \end{aligned} \tag{C.11}$$

The other six factorizations are found by interchanging the variables. If θ_2 is considered a nuisance parameter, the marginalization of Equations (C.10) and (C.11) writes

$$\begin{aligned} \int_{\theta_2} p(\theta_1, \theta_2, y)\,\mathrm{d}\theta_2 &= \int_{\theta_2} p(\theta_1|\theta_2, y)p(\theta_2|y)p(y)\,\mathrm{d}\theta_2 \\ &= \int_{\theta_2} p(\theta_1, \theta_2|y)p(y)\,\mathrm{d}\theta_2, \end{aligned} \tag{C.12}$$

which yields

$$p(\theta_1, y) = p(\theta_1|y)p(y). \tag{C.13}$$

From Equation (C.12) it can be seen that

$$\int_{\theta_2} p(\theta_1|\theta_2, y)p(\theta_2|y)\,\mathrm{d}\theta_2 = \int_{\theta_2} p(\theta_1, \theta_2|y)\,\mathrm{d}\theta_2, \tag{C.14}$$

which is a useful expression for marginalization. When the marginalization of θ_2 is applied to Bayes' Theorem, with y being observed, which may be written as

$$\int_{\theta_2} p(\theta_1, \theta_2|y)\,\mathrm{d}\theta_2 = \frac{\int_{\theta_2} p(y|\theta_1, \theta_2)p(\theta_1, \theta_2)\,\mathrm{d}\theta_2}{p(y)}, \tag{C.15}$$

this yields

$$p(\theta_1|y) = \frac{p(y|\theta_1)p(\theta_1)}{p(y)}, \tag{C.16}$$

which yields a simpler expression for the posterior distribution of θ_1 given y, without the burden of an extra parameter.

C.5 Arc reversal

As mentioned before, a Bayesian graph is a Directed Acyclic Graph. This means that the nodes in the graph are connected by arrows showing the dependencies between the nodes, and that from any node in the network there is no directed path back to itself. For making inferences in a Bayesian graph, it can be useful to change the directions of several arcs, the so called arc reversal. Arc reversal [*Shachter*, 1986; *Shachter*, 1988; *Cinicioglu and Shenoy*, 2009; *Kjærulff and Madsen*, 2012, p. 56,116] is a technique to change the dependencies between RVs in a Bayesian graph, but without changing the joint distribution of the total graph. The method is based on multiple application of Bayes' Theorem [*Shachter*, 1986; *Kjærulff and Madsen*, 2012, p. 56,116], which subsequently is a multiple factorization of the joint distribution [*Gelman et al.*, 2014, p. 63]. Although the joint distribution of the network does not change, the probability functions, or potentials, of each node mostly will change. Furthermore, the direction of the arc means probabilistic dependency and does not necessarily represent causality.

When a node of a Bayesian graph is observed, the probability function of the network changes, given this observation. One may be interested in the probability function of a certain node, the node of interest, given the observation. This probability function is the posterior distribution function of that specific node. The aim of arc reversal is to pass the information of an observation, also called a message, through the network to obtain the posterior distribution of the node of interest [*Bishop*, 2006, p. 394]. After an arc reversal, the arcs in the graph are pointing from the observed node to the node of interest.

A generic example of the working of an arc reversal is shown in Figure C.8, which is proven in *Shachter* [1986, 1988] and replicated here. In the graph in Figure C.8a is arc (XY) the arc to be reversed. The nodes U and W represent all parent

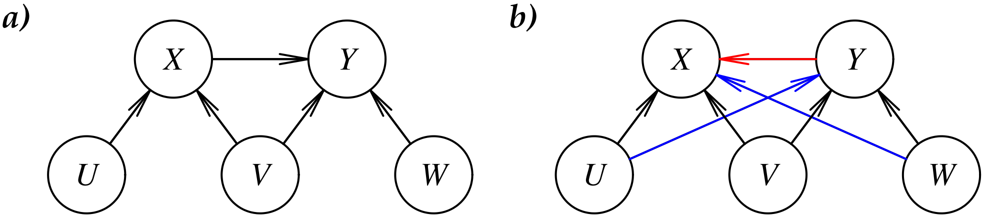

Figure C.8: Arc reversal example, pane a) shows the original graph and pane b) the graph after arc reversal. The red arc (YX) is the reversed arc and the blue arcs, (UY) and (WX), are added as a consequence of the arc reversal.

nodes of X and Y, respectively. The node V represents the common parents of both nodes X and Y. The joint distribution according to Figure C.8a can be factorized as

$$p(x, y, u, v, w) = p(y|x, v, w)p(x|u, v)p(u, v, w), \tag{C.17}$$

where $p(u, v, w) = p(u)p(v)p(w)$ because of their mutual independence in the given graph. After the arc reversal, Y is independent of X. This can be achieved by integrating out X from Equation (C.17), as shown in Equations (C.12) and (C.13). This yields

$$\begin{aligned} p(y, u, v, w) &= p(u, v, w) \int_x p(y|x, v, w)p(x|u, v)\mathrm{d}x \\ &= p(u, v, w)p(y|u, v, w), \end{aligned} \tag{C.18}$$

where $p(y|u, v, w)$ is the new distribution function, or potential, of Y. Herewith Y depends on U too, so arc (UY) has to be added to the graph (Figure C.8b). The factorization of the joint distribution after the arc reversal can be written as

$$p(x, y, u, v, w) = p(y|u, v, w)p(x|y, u, v, w)p(u, v, w), \tag{C.19}$$

with $p(y|u, v, w)$ being defined through Equation (C.18), but with $p(x|y, u, v, w)$ currently unknown. Equating Equations (C.17) and (C.19) yields

$$p(y|u, v, w)p(x|y, u, v, w)p(u, v, w) = p(y|x, v, w)p(x|u, v)p(u, v, w), \tag{C.20}$$

which, by application of Bayes' rule, can be rewritten as

$$p(x|y, u, v, w) = \frac{p(y|x, v, w)p(x|u, v)}{p(y|u, v, w)}. \tag{C.21}$$

From the right hand side of this equation it is clear that the conditional probability function of x depends on w as well now, which is denoted by the blue arc (WX) in Figure C.8.

D

Likelihood marginalization with uncertain observations

In Section 5.1.6, an expression is given to evaluate the likelihood function using uncertain observations. Here, the derivation of the marginalization is given.

To find the marginal distributions of any parameter θ_i, equation

$$\ell(\theta|X_{1,\dots,\nu}) = \nu \int_X \bar{f}(x) \ln\left(f(x|\theta)\right) \mathrm{d}x. \tag{D.1}$$

which is Equation (5.55), needs to be integrated over X to find the likelihoods of θ. Both functions, $\bar{f}(x)$ and $f(x;\theta)$ are described by piecewise linear density functions. The two piecewise linear functions are, for interval $x \in [a, b]$ and a fixed value of θ, defined as

$$\bar{f}_{ab}(x) = r(x-a) + p, \tag{D.2}$$

with $r = (\bar{f}(b) - \bar{f}(a))/(b-a)$ and $p = \bar{f}(a)$, and

$$f_{ab}(x;\theta) = s(x-a) + q, \tag{D.3}$$

with $s = (f(b;\theta) - f(a;\theta))/(b-a)$ and $q = f(a;\theta)$. So the integral with respect to x for one bin of $x \in [a, b]$ yields

$$\ell_{ab}(\theta|X_{1,\dots,\nu}) = \nu \int_{x=a}^{b} (r(x-a)+p) \ln\left(s(x-a)+q\right) \mathrm{d}x, \tag{D.4}$$

and for $t = x - a$ and $\mathrm{d}t = \mathrm{d}x$

$$\ell_{ab}(\theta|X_{1,\dots,\nu}) = \nu \int_{t=0}^{b-a} (rt+p) \ln\left(st+q\right) \mathrm{d}t, \tag{D.5}$$

and further for $u = st + q$ and $\mathrm{d}u = s\,\mathrm{d}u$

$$\begin{aligned} \ell_{ab}(\theta|X_{1,\dots,\nu}) &= \nu \int_{u=q}^{s(b-a)+q} (r(u-q)/s + p) \ln\left(u\right) s\,\mathrm{d}u \\ &= \nu r \int_{u=q}^{s(b-a)+q} u \ln\left(u\right) \mathrm{d}u + \nu(ps - rq) \int_{u=q}^{s(b-a)+q} \ln\left(u\right) \mathrm{d}u \\ &= \nu \big[r \tfrac{1}{4} u^2 (2\ln(u) - 1) + (ps - rq) u (\ln(u) - 1) \big]_q^{s(b-a)+q}. \end{aligned} \tag{D.6}$$

From Equation (D.3) it can be seen that the integration boundaries of u are $q = f_{ab}(a;\theta)$ and $s(b-a) + q = f_{ab}(b;\theta)$. For $u = 0$ the expression between the brackets is 0, since $\lim_{u\downarrow 0} u \ln(u) = 0$. The log-likelihood now writes

$$\ell(\theta|X_{1,\dots,\nu}) = \sum_{j=1}^{n_f} \ell_j(\theta|X_{1,\dots,\nu}), \tag{D.7}$$

where n_f is the number of bins in the piecewise linear function f.

The posterior distribution of θ_i can now be found by integrating out all the parameters except θ_i. The parameter θ_i stands for any of the marginal parameters P_{K_i}, S_{K_i}, X_{K_i}, P_{D_i}, S_{D_i} or X_{D_i}. The last step is integrating out of the marginal distributions of K_i and D_i, written in arithmetic form as $K_i = P_{K_i} + S_{K_i} X_{K_i}$ and $D_i = P_{D_i} + S_{D_i} X_{D_i}$. This yields the posterior predictive distributions of K_i and D_i, which is the aim of all the effort.

Samenvatting

EEN MODEL is een poging om de werkelijkheid te begrijpen en te beschrijven. Het doel van een model is om uitspraken over eigenschappen van de werkelijkheid te kunnen doen zonder dat die gemeten zijn of redelijkerwijs gemeten kunnen worden. Een voorbeeld hiervan is een model om het weer van de komende dagen te voorspellen.

Ieder model beschrijft de werkelijkheid in meer of mindere mate, maar zal nooit exact met de werkelijkheid overeen komen. Of een model bruikbaar is hangt af van de toepassing daarvan. Als je wilt weten of het morgen regent dan is het voldoende om te weten dat er met 99,8% zekerheid tussen de 5 en de 95 mm neerslag zal vallen; het zal regenen. Als iemand wil weten of het riool morgen de neerslag kan verwerken dan is deze weersverwachting onvoldoende. Voor deze laatste toepassing zal het weermodel een verbeteringsslag moeten ondergaan.

Onderzoeksvraag

In dit proefschrift wordt gebruik gemaakt van twee soorten modellen: grondwater(stromings)modellen en hydrogeologische modellen. Hierbij is het laatste model vaak een onderdeel van het eerste. Grondwatermodellen worden gemaakt om uitspraken te kunnen doen over, hoe kan het anders, grondwater. In het kader van dit proefschrift worden met 'grondwatermodellen' computermodellen bedoeld waarmee de stroming en de stijghoogte (het grondwaterniveau) van het grondwater berekend kunnen worden. De grondwaterstroming wordt door veel invloeden bepaald zoals neerslag en verdamping, oppervlaktewater en grondwaterwinning. Afhankelijk van het doel van het grondwatermodel zal deze data daarin gemodelleerd moeten worden. De basis van een grondwatermodel bestaat veelal uit de beschrijving van de ondergrond door middel van een hydrogeologisch model. Een dergelijk model beschrijft de samenstelling van de ondergrond en de bijbehorende doorlatendheden voor grondwaterstroming. In de Nederlandse situatie bestaat de ondergrond voornamelijk uit klei, zand en veen en mengvormen van deze sedimenten.

Een belangrijk hydrogeologisch model in Nederland is REGIS. Dit model wordt ontwikkeld en onderhouden bij de Geologische Dienst Nederland (TNO-GDN). Voor veel grondwatermodellen dient REGIS als basis voor de beschrijving van de ondergrond. Het REGIS model bestaat uit meer dan honderd lagen, wat meestal onnodig en onwerkbaar veel is voor een grondwatermodel. Afhankelijk van de toepassing van het grondwatermodel zal er een mate van vereenvoudiging plaatsvinden en zullen meerdere lagen uit het REGIS model samengevoegd worden tot één laag in het grondwatermodel. Vaak blijkt dat een grondwatermodel in eerste instantie onvoldoende nauwkeurige resultaten oplevert voor een beoogde toepassing. Daarom moet het model gecalibreerd worden. Bij een calibratie worden de uitkomsten (zoals stijghoogten en fluxen) van het model vergeleken met bekende waarden (metingen). Het grondwatermodel zal tijdens de calibratie zodanig aangepast worden dat de verschillen tussen de modeluitkomsten en de metingen acceptabel zijn. Bij de calibratie van het grondwatermodel zullen ook de parameters van het vereenvoudigde hydrogeologische model veranderen. En hiermee komen we bij de kern van dit proefschrift: het vereenvoudigde hydrogeologische model van het grondwatermodel is aangepast, en naar we aannemen verbeterd, maar het oorspronkelijke model REGIS niet. De belangrijkste onderzoeksvraag is daarmee:

> *Ontwikkel een methode of procedure waarmee het REGIS model kan profiteren van de verbeteringen in het grondwatermodel.*

Oftewel, bedenk een terugkoppeling.

De onderzoeksvraag is benaderd vanuit de wetenschap dat alle data een zekere mate van onzekerheid kent (stochastisch is). Als bijvoorbeeld een grondwaterstand in een peilbuis gemeten is ten opzichte van de bovenkant van de buis tot op de centimeter nauwkeurig, dan ligt de echte waarde waarschijnlijk in het interval plus of min 5 mm. Als vervolgens de hoogte van het meetpunt ten opzichte van NAP, de bovenkant van de buis, op een decimeter nauwkeurig bekend is, dan is de hoogte van de grondwaterstand niet meer op de centimeter maar op de decimeter nauwkeurig bekend. Deze onzekerheid geldt, vaak zelfs in sterkere mate, ook voor de doorlatendheden en laagdiktes van de lagen in het hydrogeologische model. Deze laatste twee parameters, de laagdikte en de doorlatendheid, zijn van belang in de voorliggende studie. Dergelijke parameters waarvan de waarde onzeker is zijn zogenaamde kansvariabelen. Om de onzekerheid van de kansvariabelen te kunnen kwantificeren wordt gebruik gemaakt van kansverdelingen.

Onzekerheid

Bij het kwantificeren van onzekerheid wordt vaak gebruikt gemaakt van standaard kansverdelingen, zoals Gaussische of log-normale verdelingen. Dit heeft onder andere als voordeel dat bepaalde rekenkundige bewerkingen in relatief weinig rekentijd uitgevoerd kunnen worden en dat de uitkomst een volledig bepaalde kansverdeling is. Het aantal bewerkingen waar dit voor geldt is echter beperkt, niet elke bewerking levert een standaardvorm kansverdeling op. Om een hoge mate van flexibiliteit in de kwantificering van de onzekerheid in parameters en de vrijheid in de rekenkundige bewerkingen te hebben, is gekozen om alle kansverdelingen te beschrijven met behulp van gelineariseerde kansdichtheidsfuncties. Meestal wordt een kansdichtheidsfunctie beschreven met behulp van een analytische functie, de curve hiervan is vaak een vloeiende lijn over het hele domein van de functie. In principe bestaat een dergelijke curve uit oneindig veel punten. Bij een gelineariseerde kansdichtheidsfunctie is een beperkt aantal punten (tientallen) op de curve gekozen die verbonden zijn door rechte lijnstukken. Deze discretisatiepunten zijn zodanig gekozen dat de fout die dit veroorzaakt minimaal is. Alle benodigde bewerkingen voor dit onderzoek zijn uitgewerkt voor gelineariseerde kansdichtheidsfuncties.

Terug naar de doelstelling van dit onderzoek: is het mogelijk om met behulp van de gecalibreerde hydrogeologische beschrijving van een grondwatermodel het REGIS model te verbeteren? Daarnaast is de vraag: kan er rekening gehouden worden met onzekerheden in de parameterwaarden? Dat wil zeggen, laat gegevens met een grotere onzekerheid een lager gewicht krijgen in de terugkoppeling dan gegevens met een kleinere onzekerheid. Doordat REGIS als basis dient voor meerdere grondwatermodellen kunnen er meerdere gecalibreerde grondwatermodellen beschikbaar zijn op dezelfde locatie. Daarmee komen we bij de laatste vraag: is het mogelijk om al deze modellen te gebruiken om REGIS te verbeteren? (Spoiler: ja.)

Uitwerking van de vraagstelling

De beantwoording van de onderzoeksvragen is in verschillende stappen gedaan. De eerste stap is de ontwikkeling van berekeningsmethoden met gelineariseerde kansdichtheidsfuncties. Bij het type berekeningen dat ondersteund moet worden, moet gedacht worden aan rekenkundige bewerkingen zoals optellen en vermenigvuldigen van kansvariabelen, toepassen van functies zoals logaritme en worteltrekken en het zoeken naar maximum waarden in kansverdelingen onder bepaalde voorwaarden. Voor gelineariseerde kansverdelingen is dit geen gemeengoed in de stochastiek en het was daarom noodzakelijk om daar algoritmes en programmatuur voor te ontwikkelen. In hoofdstuk 2 is een groot deel van deze bewerkingen beschreven en is de werking hiervan aangetoond op een ruimtelijke interpolatie (kriging interpolatie) van ondergrondgegevens. Bij de vervaardiging van een hydrogeologisch model, zoals REGIS, zijn metingen aan de ondergrond uitgevoerd met behulp van boringen. Hiermee is op een bepaald punt (of beter gezegd, een verticale lijn) de samenstelling van de ondergrond bekend. Afhankelijk van het type boring zijn de eigenschappen van de ondergrond in meer of mindere mate met zekerheid bekend. Op de locaties waar niet geboord is, is de samenstelling van de ondergrond onbekend. Door middel van interpolatie kan op de onbemeten locaties toch iets gezegd worden over de sedimenten die daar aanwezig zijn, maar de onzekerheid hierover is groter dan op de bemeten locaties. Voor laagdiktes en (hydraulische) doorlatendheden is deze interpolatie uitgevoerd waarbij de onzekerheden in de metingen meegenomen kunnen worden. Hierdoor ontstaan vlakdekkende kaarten waarbij op elke locatie een kansverdeling van de laagdikte en de doorlatendheid bekend is.

Na deze eerste stap zijn de laagdiktes en doorlatendheden van een hydrogeologisch model (in dit geval REGIS) bekend, inclusief een beschrijving van de onzekerheid. Deze parameterwaarden zijn ruimtelijk verschillend en zijn bepaald op een regelmatig grid met cellen van 100 m×100 m. Wanneer in hetzelfde gebied ook een gecalibreed grondwatermodel (op basis van REGIS) beschikbaar is, waarbij de ondergrond parameters zijn aangepast en naar we aannemen verbeterd zijn, is dat extra informatie. Een terugkoppeling van de verbeterde parameterwaarden uit het grondwatermodel naar het hydrogeologische model REGIS is dan waardevol. Een probleem hierbij is dat de één op één koppeling tussen het hydrogeologisch model en het grondwatermodel verloren is gegaan door het samenvoegen van meerdere REGIS lagen tot één laag in het grondwatermodel. In de tweede stap (hoofdstuk 3) worden de laagdiktes en de doorlatendheden van de REGIS modellagen, die samen overeenkomen met één laag uit het grondwatermodel, aangepast op basis van de gecalibreerde parameterwaarden uit het grondwatermodel. Omdat elke gridcel in het grondwatermodel is samengesteld uit meerdere REGIS modellagen met bijbehorende kansverdelingen, en de gecalibreerde parameter uit het grondwatermodel voor diezelfde gridcel slechts één waarde heeft, is er geen unieke oplossing maar bestaan er oneindig veel oplossingen. De beschreven methode kiest nu uit deze oneindige hoeveelheid mogelijkheden de combinatie met de hoogste waarschijnlijkheid. Deze stap wordt uitgevoerd op elke locatie (gridcel) van het gebied waar beide modellen elkaar overlappen.

De sedimenten in de ondergrond worden op basis van geologische processen en de periode van afzetting ingedeeld in verschillende eenheden (formaties). Binnen een formatie kunnen weer subeenheden onderscheiden worden. Bij de bouw van een hydrogeologisch model wordt dit onder andere gedaan op basis van hydraulische eigenschappen van de sedimenten. Binnen een dergelijke subeenheid wordt meestal aangenomen dat de eigenschappen (in dit geval de doorlatendheden) van het sediment overal gelijk zijn. Of beter gezegd, er zijn onvoldoende gegevens beschikbaar om ruimtelijke verschillen te kunnen beschrijven. Bij de calibratie van een grondwatermodel is het vaak noodzakelijk om een ruimtelijke variatie in de hydraulische eigenschappen van een modellaag aan te brengen om zo een model te krijgen waarvan de uitkomsten beter aansluiten bij de gemeten waarden. Bij toepassing van de terugkoppeling, zoals hiervoor beschreven, komt deze ruimtelijke variatie ook tot uitdrukking in de parameterisatie van het REGIS model. Dit is getest op een gebied in midden Nederland waar twee gescheiden gebieden met kleiige afzettingen van de Eem Formatie voor komen. Van deze twee gebieden zijn de initiële doorlatendheden van deze sedimenten gelijk gekozen. In het gecalibreerde grondwatermodel was er echter een duidelijk verschil in (relatieve) aanpassing van de verticale weerstand te zien. Na terugkoppeling van de gecalibreerde waarden was dit verschil terug te vinden als verschil in de doorlatendheden van de twee gebieden binnen de REGIS eenheden. Of deze uitkomst ook met de werkelijkheid overeenkomt moet op basis van aanvullende gegevens gevalideerd worden.

De hiervoor beschreven methode gebruikt (per gridcel) één waarde uit het gecalibreerde grondwatermodel en geeft als resultaat één waarde voor elke parameter in het REGIS model. Deze methode is overzichtelijk en relatief eenvoudig toe te passen. De methode heeft echter als nadeel dat er slechts gebruik gemaakt kan worden van één grondwatermodel en dat er geen informatie meer beschikbaar is over de onzekerheid van de paramaters. Dit laatste geldt zowel voor de gecalibreerde waarden van het grondwatermodel als voor de parameterwaarden van het REGIS model na de terugkoppeling.

De laatste methode in dit proefschrift, beschreven in hoofdstuk 5, heeft de mogelijkheid om de resultaten van meerdere gecalibreerde grondwatermodellen te gebruiken. Die modelresultaten kunnen van een eigen onzekerheid voorzien zijn, waardoor verschillende modellen naar betrouwbaarheid gewogen kunnen worden. Het resultaat van de terugkoppeling is niet een enkele waarde, zoals in de eerdere methode, maar een volledige kansverdeling. Dit heeft als groot voordeel dat het ook na een update van de parameters van het hydrogeologisch model duidelijk is hoe betrouwbaar de resultaten zijn. Deze methode maakt gebruik van Bayesiaanse statistiek, waarbij het probleem gedefinieerd is met behulp van een Bayesiaans netwerk. De basis van die techniek is dat het toevoegen van een waarneming (of meting) de oorspronkelijke (a priori) kansverdelingen van de parameters aangepast worden en dat de parameterwaarden in betrouwbaarheid toenemen. Hoe meer (onafhankelijke) metingen er beschikbaar zijn, hoe betrouwbaarder de (a posteriori) parameters zullen worden. Als meting worden de gecalibreerde waarden van een grondwatermodel gebruikt. Het is met deze techniek dus mogelijk om resultaten van meerdere grondwatermodellen te gebruiken.

In hoofdlijnen kan gesteld worden dat er in dit onderzoek twee methodes zijn ontwikkeld om een terugkoppeling te realiseren tussen gecalibreerde grondwatermodellen en een hydrogeologisch model zoals REGIS. Daarbij wordt de onzekerheid in de parameterwaarden gekwantificeerd met behulp van kansverdelingen. Voor de noodzakelijke bewerkingen zijn niet altijd analytische oplossingen beschikbaar. Daarom zijn er ondersteunende algoritmes ontwikkeld met gelineariseerde kansverdelingen om die bewerkingen mogelijk te maken.

Bibliography

Andreon, S., and B. Weaver, *Bayesian Methods for the Physical Sciences*, Springer International Publishing, doi:10.1007/978-3-319-15287-5, 2015.

Beven, K., A manifesto for the equifinality thesis, *Journal of Hydrology*, *320*(1-2), 18–36, doi:10.1016/j.jhydrol.2005.07.007, 2006.

Beven, K., and A. Binley, The future of distributed models: Model calibration and uncertainty prediction, *Hydrol. Process.*, *6*(3), 279–298, doi:10.1002/hyp.3360060305, 1992.

Bierkens, M. F. P., and H. J. T. Weerts, Block hydraulic conductivity of cross-bedded fluvial sediments, *Water Resour. Res.*, *30*(10), 2665–2678, doi:10.1029/94WR01049, 1994.

Bishop, C. M., *Pattern Recognition and Machine Learning*, Springer-Verlag New York Inc., 2006.

Blindow, N., D. Eisenburger, B. Illich, H. Petzold, and T. Richter, Ground penetrating radar, in *Environmental Geology*, pp. 283–335, Springer Berlin Heidelberg, doi:10.1007/978-3-540-74671-3_10, 2007.

Bohrnstedt, G. W., and A. S. Goldberger, On the exact covariance of products of random variables, *J. Am. Stat. Assoc.*, *64*(328), 1439–1442, doi:10.2307/2286081, 1969.

Bolstad, W. M., *Introduction to Bayesian Statistics, 2nd Edition*, Wiley-Interscience, Hoboken, New Jersey, 2007.

Bosch, J. H. A., H. J. T. Weerts, and F. S. Busschers, Formatie van Urk. in: Lithostratigrafische Nomenclator van de Ondiepe Ondergrond, 2003a.

Bosch, J. H. A., F. S. Busschers, and H. J. T. Weerts, Eem Formatie. in: Lithostratigrafische Nomenclator van de Ondiepe Ondergrond, 2003b.

Boschan, A., and B. Nœtinger, Scale dependence of effective hydraulic conductivity distributions in 3d heterogeneous media: A numerical study, *Transport in Porous Media*, *94*(1), 101–121, doi:10.1007/s11242-012-9991-2, 2012.

Busschers, F. S., and H. J. Weerts, Formatie van Kreftenheye. in: Lithostratigrafische Nomenclator van de Ondiepe Ondergrond, 2003.

Busschers, F. S., C. Kasse, R. T. van Balen, J. Vandenberghe, K. M. Cohen, H. J. T. Weerts, J. Wallinga, C. Johns, P. Cleveringa, and F. P. M. Bunnik, Late pleistocene evolution of the rhine-meuse system in the southern north sea basin: imprints of climate change, sea-level oscillation and glacio-isostacy, *Quaternary Science Reviews*, *26*(25-28), 3216–3248, doi:10.1016/j.quascirev.2007.07.013, 2007.

Cabañas, R., A. Cano, M. Gómez-Olmedo, and A. L. Madsen, On spi-lazy evaluation of influence diagrams, in *Probabilistic Graphical Models*, edited by L. C. van der Gaag and A. J. Feelders, Springer International Publishing, 2014.

Carrera, J., A. Alcolea, A. Medina, J. Hidalgo, and L. J. Slooten, Inverse problem in hydrogeology, *Hydrogeol J*, *13*(1), 206–222, doi:10.1007/s10040-004-0404-7, 2005.

Cinicioglu, E. N., and P. P. Shenoy, Arc reversals in hybrid bayesian networks with deterministic variables, *International Journal of Approximate Reasoning*, *50*(5), 763–777, doi:10.1016/j.ijar.2009.02.005, 2009.

Cleveringa, P., T. Meijer, R. J. W. van Leeuwen, H. de Wolf, R. Pouwer, T. Lissenberg, and A. W. Burger, The eemian stratotype locality at amersfoort in the central Netherlands: a re-evaluation of old and new data, *Netherlands Journal of Geosciences*, *79*(2-3), 197–216, doi:10.1017/s0016774600023659, 2000.

Cobb, B. R., and P. P. Shenoy, Nonlinear deterministic relationships in bayesian networks, in *Lecture Notes in Computer Science*, pp. 27–38, Springer Berlin Heidelberg, doi:10.1007/11518655_4, 2005.

Cobb, B. R., and P. P. Shenoy, Operations for inference in continuous bayesian networks with linear deterministic variables, *International Journal of Approximate Reasoning*, *42*(1-2), 21–36, doi:10.1016/j.ijar.2005.10.002, 2006.

Cobb, B. R., and P. P. Shenoy, Inference in hybrid bayesian networks with nonlinear deterministic conditionals, *International Journal of Intelligent Systems*, doi:10.1002/int.21897, 2017.

Cushman, J. H., L. S. Bennethum, and B. X. Hu, A primer on upscaling tools for porous media, *Adv Water Resour*, *25*(8-12), 1043–1067, doi:10.1016/S0309-1708(02)00047-7, 2002.

Dagan, G., Statistical theory of groundwater flow and transport: Pore to laboratory, laboratory to formation, and formation to regional scale., *Water Resour. Res.*, *22*(9S), 120S–134S, doi:10.1029/WR022i09Sp0120S, 1986.

de Gans, W., D. J. Beets, and M. C. Centineo, Late saalian and eemian deposits in the amsterdam glacial basin, *Netherlands Journal of Geosciences*, *79*(2-3), 147–160, doi:10.1017/s0016774600021685, 2000.

de Lange, W., and W. Borren, Grondwatermodel azure versie 1.0, *Tech. rep.*, Deltares, The Netherlands, 2014.

De Wit, A., Correlation structure dependence of the effective permeability of heterogeneous porous media., *Phys. Fluids*, *7*(11), 2553, doi:10.1063/1.868705, 1995.

Dempster, A. P., New methods for reasoning towards posterior distributions based on sample data, *The Annals of Mathematical Statistics*, *37*(2), 355–374, 1966.

Dempster, A. P., Upper and lower probabilities induced by a multivalued mapping, *The Annals of Mathematical Statistics*, *38*(2), 325–339, doi:10.1214/aoms/1177698950, 1967.

Denœux, T., Maximum likelihood estimation from uncertain data in the belief function framework, *IEEE Trans. Knowledge Data Eng.*, *25*(1), 119–130, doi:10.1109/tkde.2011.201, 2013.

Denœux, T., Likelihood-based belief function: Justification and some extensions to low-quality data, *International Journal of Approximate Reasoning*, *55*(7), 1535–1547, doi:10.1016/j.ijar.2013.06.007, 2014.

Deutsch, C. V., Correcting for negative weights in ordinary kriging, *Comput Geosci*, *22*(7), 765 – 773, doi:10.1016/0098-3004(96)00005-2, 1996.

Efstratiadis, A., and D. Koutsoyiannis, One decade of multi-objective calibration approaches in hydrological modelling: a review, *Hydrological Sc. J.*, *55*(1), 58–78, doi:10.1080/02626660903526292, 2010.

Fiori, A., G. Dagan, and I. Jankovic, Upscaling of steady flow in three-dimensional highly heterogeneous formations, *Multiscale Model. Simul.*, *9*(3), 1162–1180, doi: 10.1137/110820294, 2011.

Gelman, A., C. J. B., H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*, third ed., CRC Press, Taylor & Francis Group, 6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742, 2014.

Goovaerts, P., *Geostatistics for natural resources evaluation*, 483 pp., Oxford University Press, New York, NY, USA, 1997.

Gunnink, J. L., D. Maljers, S. F. Van Gessel, A. Menkovic, and H. J. Hummelman, Digital geological model (dgm): A 3d raster model of the subsurface of the netherlands, *Geologie en Mijnbouw/Netherlands Journal of Geosciences*, *92*(1), 33–46, doi:10.1017/S0016774600000263, 2013.

Gupta, H. V., S. Sorooshian, and P. O. Yapo, Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information, *Water Resources Research*, *34*(4), 751–763, doi:10.1029/97wr03495, 1998.

Held, L., and D. S. Bové, *Applied Statistical Inference: Likelihood and Bayes*, Springer, Heidelberg, Germany, doi:10.1007/978-3-642-37887-4, 2013.

Hendricks Franssen, H. J., A. Alcolea, M. Riva, M. Bakr, N. van der Wiel, F. Stauffer, and A. Guadagnini, A comparison of seven methods for the inverse modelling of groundwater flow. application to the characterisation of well catchments, *Adv Water Resour*, *32*(6), 851 – 872, doi:10.1016/j.advwatres.2009.02.011, 2009.

Holmes, D. T., and K. A. Buhr, Error propagation in calculated ratios., *Clin. Biochem.*, *40*(9 - 10), 728 – 734, doi:10.1016/j.clinbiochem.2006.12.014, 2007.

Hoteit, I., X. Luo, and D.-T. Pham, Particle kalman filtering: A nonlinear bayesian framework for ensemble kalman filters, *Mon. Weather Rev.*, *140*(2), 528–542, doi: 10.1175/2011MWR3640.1, 2012.

Hristopulos, D. T., Renormalization group methods in subsurface hydrology: overview and applications in hydraulic conductivity upscaling, *Adv Water Resour*, *26*(12), 1279–1308, doi:10.1016/S0309-1708(03)00103-9, 2003.

Hristopulos, D. T., and G. Christakos, Renormalization group analysis of permeability upscaling, *Stoch Environ Res Risk Assess*, *13*(1-2), 131–160, doi:10.1007/s004770050036, 1999.

Isaaks, E. H., and R. M. Srivastava, *An Introduction to Applied Geostatistics*, 561 pp., Oxford University Press, New York, NY, USA, 1989.

Izenman, A. J., Recent developments in nonparametric density estimation, *J. Am. Stat. Assoc.*, *86*(413), 205–224, doi:10.2307/2289732, 1991.

Jaroszewicz, S., and M. Korzeń, Arithmetic operations on independent random variables: A numerical approach, *SIAM Journal on Scientific Computing*, *34*(3), A1241–A1265, doi:10.1137/110839680, 2012.

Jaynes, E. T., *Probability Theory*, Cambridge University Press, doi:10.1017/CBO9780511790423.019, 2003.

Journel, A. G., and C. J. Huijbregts, *Mining Geostatistics*, fifth printing 1991 ed., The Blackburn Press, Caldwell, New Jersey, USA, 1978.

Kaczynski, W., L. Leemis, N. Loehr, and J. McQueston, Nonparametric random variate generation using a piecewise-linear cumulative distribution function, *Commun Stat Simul Comput*, *41*(4), 449–468, doi:10.1080/03610918.2011.606947, 2012.

Khuri, A. I., Applications of dirac's delta function in statistics, *International Journal of Mathematical Education in Science and Technology*, *35*(2), 185–195, doi:10.1080/00207390310001638313, 2004.

King, P. R., The use of renormalization for calculating effective permeability, *Transport in Porous Media*, *4*(1), 37–58, doi:10.1007/BF00134741, 1989.

Kjærulff, U. B., and A. L. Madsen, *Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis*, Springer New York, 2012.

Korzeń, M., and S. Jaroszewicz, Pacal: A python package for arithmetic computations with random variables, *Journal of Statistical Software*, *57*(10), 1–34, doi:10.18637/jss.v057.i10, 2014.

Kozlov, A. V., and D. Koller, Nonuniform dynamic discretization in hybrid networks, in *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, UAI'97, pp. 314–325, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

Kroese, D. P., T. Taimre, and Z. I. Botev, *Handbook of Monte Carlo Methods*, John Wiley and Sons Ltd, 2011.

Kyriakidis, P., and P. Gaganis, Efficient simulation of (log)normal random fields for hydrogeological applications, *Math. Geosci.*, *45*(5), 531–556, doi:10.1007/s11004-013-9470-5, 2013.

Lang, F. D. d., and H. J. T. Weerts, Formatie van Stramproy. in: Lithostratigrafische Nomenclator van de Ondiepe Ondergrond, 2003.

Lange, W. J. D., G. F. Prinsen, J. C. Hoogewoud, A. A. Veldhuizen, J. Verkaik, G. H. P. O. Essink, P. E. V. van Walsum, J. R. Delsman, J. C. Hunink, H. T. L. Massop, and T. Kroon, An operational, multi-scale, multi-model system for consensus-based, integrated water management and policy analysis: The Netherlands hydrological instrument, *Environmental Modelling & Software*, *59*, 98–108, doi:10.1016/j.envsoft.2014.05.009, 2014.

Langseth, H., T. D. Nielsen, R. Rumí, and A. Salmerón, Maximum likelihood learning of conditional MTE distributions, in *Lecture Notes in Computer Science*, pp.

240–251, Springer Berlin Heidelberg, doi:10.1007/978-3-642-02906-6_22, 2009.

Long, A. J., N. L. M. Barlow, F. S. Busschers, K. M. Cohen, W. R. Gehrels, and L. M. Wake, Near-field sea-level variability in northwest Europe and ice sheet stability during the last interglacial, *Quaternary Science Reviews*, *126*, 26–40, doi:10.1016/j.quascirev.2015.08.021, 2015.

Lourens, A., and F. C. van Geer, Uncertainty propagation of arbitrary probability density functions applied to upscaling of transmissivities, *Stoch Environ Res Risk Assess*, *30*(1), 237–249, doi:10.1007/s00477-015-1075-8, 2016.

Marquez, D., M. Neil, and N. Fenton, Improved reliability modeling using bayesian networks and dynamic discretization, *Reliability Engineering & System Safety*, *95*(4), 412–425, doi:10.1016/j.ress.2009.11.012, 2010.

Marsily, G. D., *Quantitative Hydrogeology*, Academic Press, 1986.

McDonald, M. G., and A. W. Harbaugh, *A modular three-dimensional finite-difference ground-water flow model*, vol. book 6, Chapter A1, 586 pp., U.S. Geological Survey, 1988.

Meija, J., Solution to random error propagation challenge., *Anal. Bioanal. Chem.*, *396*(1), 187–188, doi:10.1007/s00216-009-3255-1, 2010.

Neil, M., M. Tailor, and D. Marquez, Inference in hybrid bayesian networks using dynamic discretization, *Statistics and Computing*, *17*(3), 219–233, doi:10.1007/s11222-007-9018-y, 2007.

Neil, M., X. Chen, and N. Fenton, Optimizing the calculation of conditional probability tables in hybrid bayesian networks using binary factorization, *IEEE Transactions on Knowledge and Data Engineering*, *24*(7), 1306–1312, doi:10.1109/tkde.2011.87, 2012.

Nœtinger, B., V. Artus, and G. Zargar, The future of stochastic and upscaling methods in hydrogeology, *Hydrogeol J*, *13*, 184–201, doi:10.1007/s10040-004-0427-0, 2005.

Pancaldi, V., K. Christensen, and P. R. King, Permeability up-scaling using haar wavelets, *Transport in Porous Media*, *67*(3), 395–412, doi:10.1007/s11242-006-9032-0, 2007.

Papoulis, A., *Probability, random variables, and stochastic processes*, McGraw-Hill electrical and electronic engineering series, McGraw-Hill, 1991.

Papoulis, A., and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill Education Ltd, 2002.

Pearl, J., Fusion, propagation, and structuring in belief networks, *Artificial Intelligence*, *29*(3), 241–288, doi:10.1016/0004-3702(86)90072-x, 1986.

Pearl, J., Belief networks revisited, *Artificial Intelligence*, *59*(1-2), 49–56, doi:10.1016/0004-3702(93)90169-c, 1993.

Peeters, J., F. S. Busschers, and E. Stouthamer, Fluvial evolution of the rhine during the last interglacial-glacial cycle in the southern north sea basin: A review and look forward, *Quaternary International*, *357*, 176–188, doi:10.1016/j.quaint.2014.03.024, 2015.

Peeters, J., F. Busschers, E. Stouthamer, J. Bosch, M. V. den Berg, J. Wallinga, A. Versendaal, F. Bunnik, and H. Middelkoop, Sedimentary architecture and chronostratigraphy of a late quaternary incised-valley fill: A case study of the late middle and late pleistocene rhine system in the Netherlands, *Quaternary Science Reviews*, *131*, 211–236, doi:10.1016/j.quascirev.2015.10.015, 2016.

Rumí, R., and A. Salmerón, Approximate probability propagation with mixtures of truncated exponentials, *International Journal of Approximate Reasoning*, *45*(2), 191–210, doi:10.1016/j.ijar.2006.06.007, 2007.

Sanchez-Vila, X., A. Guadagnini, and J. Carrera, Representative hydraulic conductivities in saturated groundwater flow, *Rev. Geophys.*, *44*(3), 1–46, doi:10.1029/2005RG000169, 2006.

Schuck, A., and G. Lange, Seismic methods, in *Environmental Geology*, pp. 337–402, Springer Berlin Heidelberg, doi:10.1007/978-3-540-74671-3_11, 2007.

Shachter, R. D., Evaluating influence diagrams, *Operations Research*, *34*(6), 871–882, 1986.

Shachter, R. D., Probabilistic inference and influence diagrams, *Operations Research*, *36*(4), 589–604, doi:10.1287/opre.36.4.589, 1988.

Shafer, G., *A Mathematical Theory of Evidence*, PRINCETON UNIV PR, 1976.

Shenoy, P. P., and J. C. West, Inference in hybrid bayesian networks using mixtures of polynomials, *International Journal of Approximate Reasoning*, *52*(5), 641–657, doi:10.1016/j.ijar.2010.09.003, 2011.

Silverman, M. P., W. Strange, and T. C. Lipscombe, The distribution of composite measurements: How to be certain of the uncertainties in what we measure, *Am. J. Phys.*, *72*(8), 1068–1081, doi:10.1119/1.1738426, 2004a.

Singh, A., B. S. Minsker, and A. J. Valocchi, An interactive multi-objective optimization framework for groundwater inverse modeling, *Advances in Water Resources*, *31*(10), 1269–1283, doi:10.1016/j.advwatres.2008.05.005, 2008.

Snepvangers, J., B. Minnema, W. Berendrecht, P. Vermeulen, A. Lourens, W. Van Der Linden, M. Duijn, J. Van Bakel, W.-J. Zaadnoordijk, M. Boerefijn, M. Meeuwissen, and V. Lagendijk, Mipwa: Water managers develop their own high-resolution groundwater model tools, in *IAHS-AISH Publication*, pp. 108–113, 2008.

Stafleu, J., D. Maljers, J. L. Gunnink, A. Menkovic, and F. S. Busschers, 3d modelling of the shallow subsurface of zeeland, the netherlands, *Netherlands Journal of Geosciences - Geologie en Mijnbouw*, *90*(4), 293–310, doi:10.1017/s0016774600000597, 2011.

TNO-GSN, DINOLoket Data and Information on the Dutch Subsurface, url: http://www.dinoloket.nl/en, Accessed: 2021-03-05, 2021.

Tolosana-Delgado, R., and V. Pawlowsky-Glahn, Kriging regionalized positive variables revisited: Sample space and scale considerations, *Math. Geol.*, *39*(6), 529–558, doi:10.1007/s11004-007-9107-7, 2007.

Tran, T., The 'missing scale' and direct simulation of block effective properties, *J.*

Hydrol., *183*(1-2), 37–56, doi:10.1016/S0022-1694(96)80033-3, 1996.

Valstar, J. R., D. B. McLaughlin, C. B. M. te Stroet, and F. C. van Geer, A representer-based inverse method for groundwater flow and transport applications, *Water Resour. Res.*, *40*(5), W05,116, doi:10.1029/2003WR002922, 2004.

Vander Wielen, M. J., and R. J. Vander Wielen, The general segmented distribution, *Commun Stat Theory Methods*, *44*, 1994, doi:10.1080/03610926.2012.758743, 2015.

Vernes, R. W., and T. H. M. van Doorn, REGIS II, a 3D hydrogeological model of The Netherlands, in *Geological Society of America, Abstracts with Programs*, vol. 38, p. 109, The Geological Society of America, Philadelphia, PA, USA, proceedings of the Philadelphia annual meeting of The Geological Society of America, 2006.

Vernes, R. W., T. H. M. van Doorn, M. F. P. Bierkens, S. F. van Gessel, and E. de Heer, Van gidslaag naar hydrogeologisch eenheid, toelichting op de totstandkoming van de dataset REGIS II (in Dutch), *Tech. rep.*, Nederlands Instituut voor Toegepaste Geowetenschappen TNO - Geological Survey of the Netherlands, Utrecht, the Netherlands, 2005.

Westerhoff, W. E., Formatie van Sterksel. in: Lithostratigrafische Nomenclator van de Ondiepe Ondergrond, 2003.

Zagwijn, W. H., Sea-level changes in The Netherlands during the Eemian, *Geologie en Mijnbouw*, *62*(3), 437–450, 1983.

Zhu, J., and M. Collette, A dynamic discretization method for reliability inference in dynamic bayesian networks, *Reliability Engineering & System Safety*, *138*, 242–252, doi:10.1016/j.ress.2015.01.017, 2015.

Zimmerman, D. A., G. de Marsily, C. A. Gotway, M. G. Marietta, C. L. Axness, R. L. Beauheim, R. L. Bras, J. Carrera, G. Dagan, P. B. Davies, D. P. Gallegos, A. Galli, J. Gómez-Hernández, P. Grindrod, A. L. Gutjahr, P. K. Kitanidis, A. M. Lavenue, D. McLaughlin, S. P. Neuman, B. S. RamaRao, C. Ravenne, and Y. Rubin, A comparison of seven geostatistically based inverse approaches to estimate transmissivities for modeling advective transport by groundwater flow, *Water Resour. Res.*, *34*(6), 1373–1413, doi:10.1029/98WR00003, 1998.

About the Author

ARIS LOURENS was born on the 21st of January in 1962 in Aalten. Aris attended secondary school at the Christelijke Scholengemeenschap Aalten, where he finished HAVO in 1981. After secondary school he studied land and water management at the Hogere Bosbouw en Cultuurtechnische School in Velp (College for Forestry and Land and Water Management). In 1985, he started as geohydrologist at the groundwater department of TNO, first in Delft and later on in Utrecht. His work is mainly involved with geostatistics and time series analysis, which all have applications in quantitative groundwater hydrology. Thereby, development of software tools, which can support the research activities, is always an important part of his work. Due to a reorganization, started in 2008, involving the establishment of the research institute Deltares, his appointment moved to this institute. In September 2011, Aris quit his job at Deltares and started a PhD research at the University of Utrecht at the Department of Physical Geography, with great pleasure. Before ending his PhD, he joined TNO again in 2017 at Geomodelling, a department of the Geological Survey of the Netherlands. To date, he can make a contribution to the maintenance and development of the hydrogeological an geological models of the department. These models perform in his PhD research too. In his spare time, Aris finished his PhD research.

Software

IN THE PROJECT OF THIS THESIS, software has been developed to be able to make the required calculations. One of the main developments was to make calculations with random variables, defined by piecewise linear probabilitye density functions (PL-PDFs), feasible. Part of this software is made publicly available. This includes the creation of PL-PDFs, arthmetic binary operations $(+-\times/)$, as described in Appendix A, and elementary functions like $\log()$ and $\exp()$. The software is written in R[1], with some core functionality written in FORTRAN for speedup reasons. The software is combined into an R package (named plPDF).

The PL-PDF objects are implemented as an S3 object class (plpdf), and methods for generic functions are supplied. Herewith, the dispatch mechanism of R is available, which makes the use of the functionality very intuitive.

The package is available at: `https://github.com/lourensa/plPDF`

[1] R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Utrecht University
Faculty of Geosciences
Department of Physical Geography

ISSN 2211-4335

Propositions belonging to the doctoral thesis:

Improving hydrogeological models using the results of calibrated groundwater flow models

A probabilistic approach using piecewise linear probability density functions and Bayesian networks

by Aris Lourens

1. Without uncertainty it is impossible to improve a model (this thesis).
2. The error made by describing a probability density function (PDF) by a piecewise linear approximation is often smaller than the error between the assumed PDF and the phenomena it describes (this thesis).
3. Availability of calculations with piecewise linear PDFs reduces the need for Monte Carlo solutions (this thesis).
4. A groundwater flow model is not equal to a hydrogeological model (this thesis).
5. A certain most likely result is more uncertain than an uncertain Bayesian result (this thesis).
6. Often, a problem is solved in Word or LaTeX, but a solution in FORTRAN is still missing.
7. A remarkable difference between literary and scientific literature is that in the first the frustrations of the author resonate and in the second the euphoria.
8. It would do justice to the referendum question if the ballot paper, besides the options ′for′ and ′against′, also would contain the option ′42′.
9. Every attainment is the result of a change.
10. Those who want war emphasize the differences, those who want peace the similarities.

Stellingen behorend bij het proefschrift:

Verbeteren van hydrogeologische modellen door resultaten van gecalibreerde grondwatermodellen te gebruiken

Een probalistische benadering met behulp van gelineairiseerde kansdichtheidsfuncties en Bayesiaanse netwerken

door Aris Lourens

1. Zonder onzekerheid is het niet mogelijk om een model te verbeteren (dit proefschrift).

2. De fout, die gemaakt wordt door een kansdichtheidsfunctie te benaderen door een gelineairiseerde functie, is vaak kleiner dan de fout tussen de veronderstelde kansdichtheidsfunctie en het beschreven fenomeen (dit proefschrift).

3. De beschikbaarheid van berekeningen met gelineairiseerde kansdichtheidsfuncties vermindert de behoefte aan Monte Carlo oplossingen (dit proefschrift).

4. Een grondwaterstromingsmodel is niet gelijk aan een hydrogeologisch model (dit proefschrift).

5. Een zekere meest waarschijnlijke uitkomst geeft minder zekerheid dan een onzekere Bayesiaanse uitkomst (dit proefschrift).

6. Vaak is een probleem opgelost in Word of LaTeX, maar ontbreekt de oplossing in FORTRAN nog.

7. Een opmerkelijk verschil tussen taalkundige en wetenschappelijke literatuur is dat bij de eerste de frustraties van de auteur veelal doorklinken en bij de tweede de euforie.

8. Het zou de vraagstelling van een referendum recht doen wanneer het stembiljet naast de opties 'voor' en 'tegen' ook de keuzemogelijkheid '42' zou bevatten.

9. Elke verworvenheid is het gevolg van een verandering.

10. Wie oorlog wil benadrukt de verschillen, wie vrede wil de overeenkomsten.