Hydrogeological conceptual model development and testing

by

Trine Enemark

Thesis

Submitted to Flinders University

for the degree of

Doctor of Philosophy

College of Science and Engineering

June 2020

Contents

Thesis summary

Groundwater models are widely applied in groundwater management to guide decision making. The success of groundwater management is directly dependent on a good understanding of the groundwater system. A conceptual model is a summary of our current knowledge about a groundwater system describing the dominating processes and the overall physical structure of the geology. One of the major sources of uncertainties in groundwater model predictions is the conceptual uncertainty that arises when more than one conceptual model can explain the available data. The goal of this thesis is to identify current approaches, unify scattered insights and develop a systematic methodology of hydrogeological conceptual model development and testing, which leads to an improved characterisation of conceptual uncertainty.

Conceptual model development involves formulation of hypotheses about the groundwater system functioning. These are the initial decisions in the modelling that drive the groundwater model predictions and form the basis of the uncertainty analysis. In this thesis we advocate for a systematic model development approach based on mutually exclusive hypotheses. We developed bold hypotheses about the model structure, challenging what was considered possible for the system, in order to give more transparent explanation of which model structures were considered possible.

Conceptual model testing consists of holding the developed models against data to evaluate their validity. Model testing is essential in order to gain confidence in the developed models and remove those models from the ensemble that are inconsistent with the data. We show that model testing does not have to be a time-consuming task but can happen in relatively simple forward models. We advocate for reserving as much data as possible for the model testing exercise rather than using all data for model development in order to be able to explain why no other conceptual models are plausible.

The methodology developed in this thesis is applied to the Wildman River area, Northern Territory, Australia. By acknowledging the existence of conceptual uncertainty, we increase the confidence in the water balance for the area. A second aspect of the investigation is the connectivity of sinkhole-like depressions in the area to groundwater and whether they may act as conduits of groundwater recharge.

The insights gained from this thesis enables more accessible methodology for conceptual model development and testing. By acknowledging and accounting for conceptual uncertainty, more confidence can be gained in groundwater model predictions leading to improved groundwater management.

Declaration

I certify that this thesis:

1. does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university; and
2. to the best of my knowledge and belief, does not contain any material previously published or written by another person except where due reference is made in the text.

Acknowledgement

I wish to express my sincere gratitude to my supervisors Luk Peeters, Dirk Mallants and Okke Batelaan for all their support and helpful guidance during my studies and for offering me the chance to do a PhD in the first place.

Thank you to friends, colleagues and visitors at CSIRO and Flinders University for suggestions and support and keeping my motivation high.

Thank you to the reviewers of the papers within this thesis including the reviewer's in CSIRO's internal review process.

Thank you to Ursula and Steve from DENR for providing valuable insights into Wildman River area and making the fieldtrips possible.

Thank you to Malcolm Sambridge and Andrew Valentine for giving me the opportunity to visit ANU and guidance and advise which was critical for the second paper to be published.

I also want to thank my former supervisor Peter Engesgaard for encouraging me to go to Australia.

The funding for this study by the CSIRO's Deep Earth Imaging – Future Science Platform and the Flinders University International Tuition Fee Sponsorship is gratefully acknowledged.

Thank you to my family and friends for your support and to Mikkel for going on this adventure with me. What a ride!

Chapter 1: Introduction

1.1 BACKGROUND

1.1.1 Groundwater management and groundwater models

As an accessible freshwater resource, groundwater represents an essential component for human needs and activities. Groundwater is used as a drinking water resource, for irrigation and thereby for food and fibre production, in industrial processes, to facilitate mining and coal seam gas extraction and to sustain the environment. Increased groundwater pumping raises the question of the sustainability of this resource (Giordano, 2009). Exploiting the groundwater system, can lead to negative impacts such as aquifer depletion (Terrell et al., 2002), land subsidence resulting from dewatering and depressurization of an aquifer (Dixon et al., 2006), seawater intrusion causing bores to be contaminated by saltwater (Post, 2005) and destabilization of groundwater dependent ecosystems, directly impacting biodiversity (de Graaf et al., 2019).

Groundwater models are widely applied for groundwater management to guide decision making as they serve as a simple but practical representation of the groundwater system in question. Groundwater models can simulate past and present conditions of the groundwater system as well as predicting future response to natural (e.g. climate) and anthropogenic stress (e.g. pumping) (Barnett et al., 2012). The success of groundwater management and avoidance of the above-mentioned consequences of overexploitation is directly dependent on a good understanding of the groundwater system (Konikow and Kendy, 2005).

As groundwater models are simple representations of a complex reality, their predictions are inherently uncertain. Characterisation of the predictive uncertainty provides the decision-maker the insights needed to understand the risks when it comes to groundwater management

(Middlemis et al., 2019). A good characterisation of the predictive uncertainty of a groundwater model has the potential to increase the likelihood to successfully identify suitable locations for developing a new groundwater resource (Sidiropoulos and Tolikas, 2004), or aid in the design of mine dewatering and provide more robust estimates of environmental impact of mine operations (Currell et al., 2017). A poor understanding of the risks involved with making decisions around groundwater use can, worst-case, lead to a negative impact on the groundwater quantity and/or quality and often results in considerable costs when remedial actions must be imposed. At best, a poor understanding of the risks involved may lead to underutilisation of the water resource and thereby missed opportunities for agricultural or industrial development or town water supply.

1.1.2 Uncertainties in groundwater models

Uncertainties in groundwater models are generally classified into model structure uncertainty (incomplete understanding and simplified description of modelled processes), parameter uncertainty (parameter values) and input uncertainty including scenario uncertainty (external driving forces) (Refsgaard et al., 2006; Vrugt, 2016; Walker et al., 2003). Parameter and input uncertainty can generally be characterised by varying the parameters or input values continuously in the interval considered plausible for the values under consideration. On the other hand, structural uncertainty is characterised by coming up with discrete alternative model structures. Uncertainties of the former are sometimes referred to as aleatory (random uncertainty) while uncertainties of the latter category are epistemic (arising from lack of knowledge) (Beven, 2016).

Conceptual uncertainty is that part of structural uncertainty that relates to the understanding of groundwater system functioning. The conceptual model definition is one of the first steps in any groundwater modelling exercise and precedes any effort to mathematically represent the groundwater system. The conceptual model provides the underlying assumptions about dominating processes and physical structure in the groundwater model, so “If the conceptual model is incorrect, advanced numerical model[l]ing can only lead to nicely presented garbage” (Jiao et al., 2005).

1.2 RESEARCH PROBLEM

The hydrogeological conceptual model is, in this thesis, considered to be a scientific theory about groundwater system functioning. Applying a realist’s viewpoint (Okasha, 2002), a conceptual model is regarded as an attempt to describe the underlying nature of reality, while an anti-realist’s viewpoint is that the conceptual model is an instrument that helps us make predictions of observational phenomena (Figure 1.1). Related to the anti-realist’s viewpoint is the underdetermination of the scientific theory. Conceptual uncertainty concerns the underdetermination of the conceptual model leading to an equifinality of conceptual models. This means that the available evidence can give rise to different conceptual understandings (Stanford, 2017). This holds true for the hydrogeological conceptual model. Several studies have shown, that experts presented with the same data, generally come up with different interpretations of the conceptual structure, e.g. (Bond et al., 2007; Højberg and Refsgaard, 2005; Seifert et al., 2012).

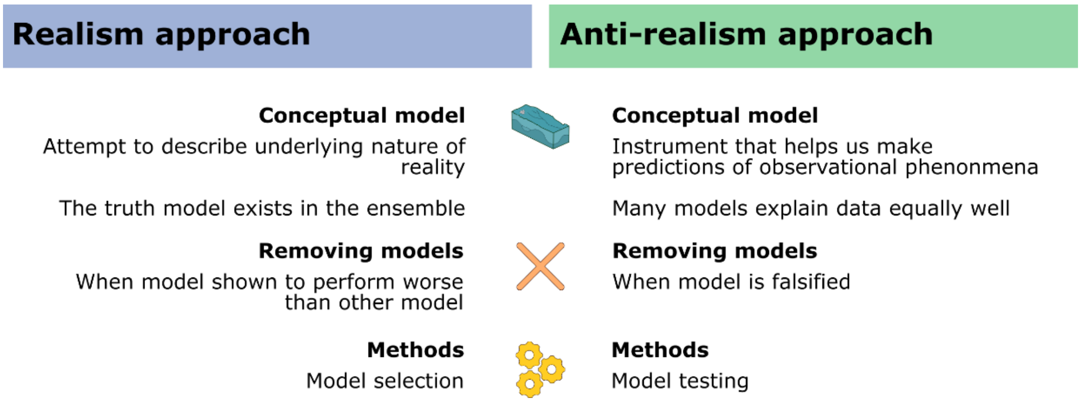
Realism approach
Anti-realism approach
Conceptual model
Attempt to describe underlying nature of reality
The truth model exists in the ensemble
Conceptual model
Instrument that helps us make predictions of observational phenonmena
Many models explain data equally well
Removing models
When model shown to perform worse than other model
Removing models
When model is falsified
Methods
Model selection
Methods
Model testing

Figure 1.1. Realism vs. anti-realism approach to conceptual modelling. The different philosophical understandings of the conceptual model lead to the application of different methods.

Traditionally in most groundwater modelling studies, only a single conceptual model is developed, thereby ignoring the conceptual uncertainty. The traditional conceptual model building approach is essentially a Kuhn cycle (Bird, 2018): a single paradigm (conceptual model) will be developed that describes the current understanding. The paradigm will be challenged by anomalies that will be explained away, but as the number and severity of anomalies built up, a crisis (or conceptual surprise) will occur. This crisis will end in a revolution that then culminates in a paradigm shift (a definition of a new conceptual model). Kuhn describes the transfer of allegiance from one paradigm to another as an act of will (based on faith and peer pressure), rather than rationality (based on evidence and reason) (Okasha, 2002). It happens only slowly and with resistance. This challenges the falsifiability of a conceptual model, which to some degree separates science from pseudo-science.

In recent years the multi-model approach has received increased attention in hydrogeology e.g. (Mustafa et al., 2020; Rojas et al., 2010c; Troldborg et al., 2007). In the multi-model approach the underdetermination of the hydrogeological conceptual model problem is acknowledged and alternative conceptual models (or paradigms) are evaluated and used to make predictions. Most attention in literature has, however, been focused on making predictions using multiple models, not on how to create a set of multiple models.

To develop a scientific theory, sometimes a distinction is made between discovery and justification. Discovery is the act of conceiving a theory, while justification is the process of justifying its claim to truth (Schickore, 2018). In hydrogeological conceptual model building discovery and justification can be correlated to conceptual model development and conceptual model testing, respectively. These are the two key aspects of the characterisation and evaluation of conceptual uncertainty. Conceptual model development involves formulation of hypotheses about the groundwater system functioning, while conceptual model testing consists of holding these hypotheses against data to evaluate their validity. In the following, these two aspects are discussed.

The objective of a groundwater model has an influence on how uncertainties should be dealt with (Middlemis et al., 2019). In the following sections we will differentiate between prediction focused and exploration focussed groundwater modelling. In prediction focused approaches, the objective is to make predictions answering management questions (Feyen and Gorelick, 2005) while in exploration focused approaches the objective is to gain a better system understanding without an immediate management question in mind (Hermans et al., 2015).

1.2.1 Model development

The objective of the model development approach in an exploration focussed exercise is to populate the model space with all plausible models, while in a prediction focussed approach it is to populate the model space with all useful models, i.e. those relevant to the prediction. When exploring the model space for all plausible models, the model development should aim at making bold hypotheses that maximizes the difference between alternative models (Guillaume et al., 2016). The models that are useful or fit-for-purpose are those of the plausible models that have a great impact on a specific prediction (Beven, 2018).

When the definition of the conceptual model is limited to the qualitative, fundamental understanding and insight into a system, theoretically it may be possible to populate the entire plausible model space, however, it is not practically possible.

No practical guidelines exist to systematically characterise conceptual uncertainty. In fact, in philosophy of science it is discussed whether the act of discovery can even be systematic (Schickore, 2018). The condition that conceptual uncertainty can only be characterised by setting of discrete alternatives rather than as a continuous range, makes it difficult to handle.

The result of the multi-model approach is generally thought to rely on the creativity of the modeller or experts (Marshall, 2017) to develop alternative models. The creativity of a modeller will always be limited by confirmation bias, that is the way we interpret data is always biased by our initial understanding. The modeller may be unaware of committing to a conceptual understanding even before the data is interpreted and the translation of conceptualization to mathematical model structure might even happen informally. Therefore the ensemble of models in a multi-model approach is thought to be a result of subjectiveness and even chance (Gondwe et al., 2010; Rajabi et al., 2018; Refsgaard et al., 2006).

As it is not practically possible to populate the plausible model space, it will consist of known unknowns and unknown unknowns. The known unknown conceptual models refer to models that we can develop, but which we are unable to discriminate between given current data. The unknown unknown conceptual models are the plausible conceptual models that we are still unaware of.

When an unknown unknown is uncovered it is often termed a conceptual surprise (Bredehoeft, 2005). However, rather it would be a surprise if we in our first attempt were successful in developing the true conceptual understanding.

A conceptual surprise related to the physical structure could be the discovery of a fault creating a barrier to groundwater flow. Another example is finding a palaeovalley that creates a direct pathway from the surface to deeper aquifers. Conceptual surprises related to the process structure could be identifying point-based groundwater as the main recharge process rather than diffuse recharge or finding evidence for a previously undetected groundwater discharge component to a lake water balance. The unknown unknowns may affect predictions such as catchment areas of wells, the vulnerability of the aquifers towards contamination and the sustainable extraction limit from aquifers. Surprises are however, rarely documented in literature due to positive publication bias (Beven, 2018), that is papers are more likely to be accepted for publication if they report successful models.

A general advice to alternative model development is that in addition to be collectively exhaustive they must be mutually exclusive (Refsgaard et al., 2012). The mutually exclusive, collectively exhaustive concepts are illustrated in Figure 1.2. The grey boxes represent the underdetermined, plausible model space that is populated with alternative models represented by the coloured boxes through the model development exercise. The alternative models are mutually exclusive when they do not overlap (representing non-overlapping ideas). They are collectively exhaustive when they cover the entire model space. The remainder of the model space that is not populated is represented by unknown unknown conceptual models. The alternative models populating the model space give rise to alternative predictions. The range of the predictions (on the x-axis) illustrates the uncertainty given the current knowledge. The aleatory uncertainty is here represented for each model by a normal distribution, but it could have any other probability distribution. The conceptual uncertainty is represented by the different models illustrated with different colours.

In Figure 1.2a a mutually exclusive and collectively exhaustive range of models has been developed. This situation is not realistic as there will always be valid conceptual understandings that we are just not aware of yet. Even if we were able to define a collectively exhaustive range of models, it would be difficult to prove that no other plausible model existed.

When the range of models is not collectively exhaustive (Figure 1.2b), that is all plausible models have not been developed, the resulting predictions are underestimating the conceptual uncertainty. As illustrated in Figure 1.2b, the predictions may also be biased. In groundwater modelling it is generally understood that not considering the uncertainties on predictions only provides the illusion of certainty (Pappenberger and Beven, 2006; Pielke, 2001).

When models are not mutually exclusive (Figure 1.2c), in that they represent similar ideas, they give rise to similar predictions. When predictions are based on many non-mutually exclusive models, a false confidence in these predictions can arise because they are based on many models. The confidence is false as one may expect more of the model space has been characterised by the many models, but as is illustrated in Figure 1.2c, this may not be the case.

In summary, a systematic model development process is important because the modeller defines the set of plausible conceptual models that form the basis for the uncertainty analysis. These are the initial decisions in the modelling that drives where the predictions will end up. For modelling studies to be reliable, a clear path from data to results must be mapped. This implies that any assumption made in the modelling process must be justified by either data or a clear identification of the reasoning behind the assumptions. This leads to increased transparency in reporting which makes the work more defensible. While this framework is generally applied, most groundwater modelling studies fail to explain why no other conceptualisations are plausible.

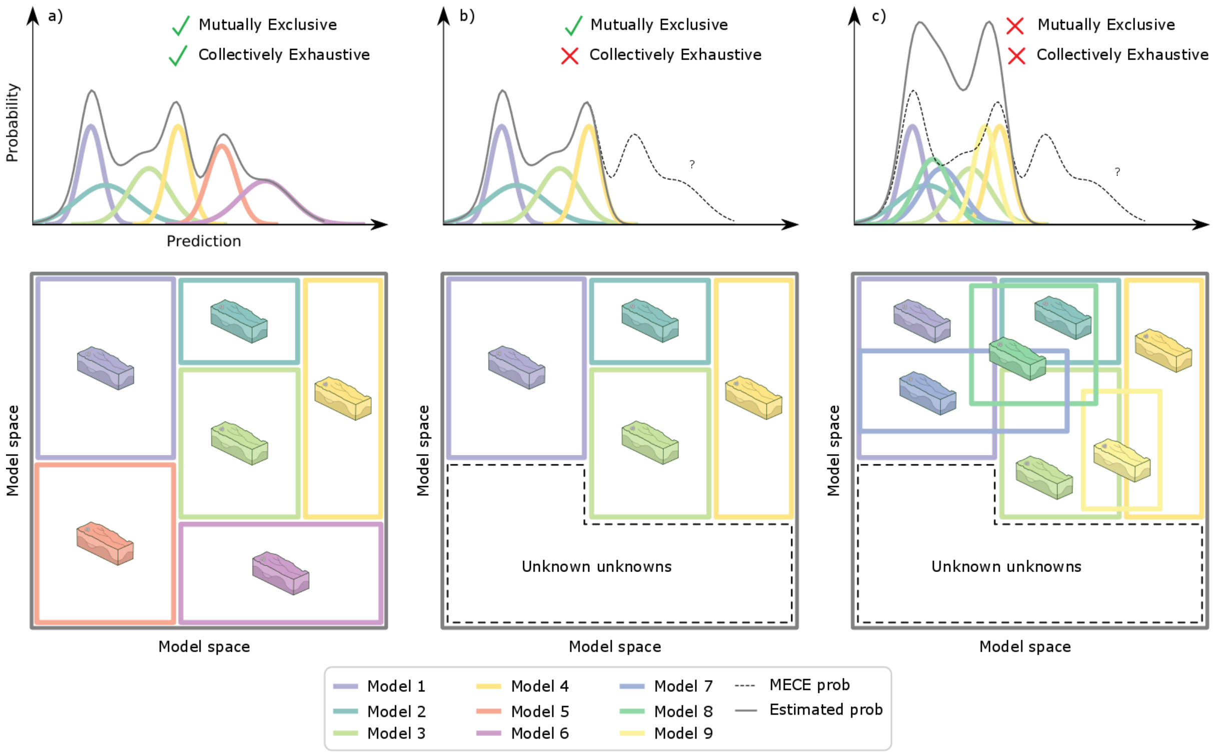

Figure 1.2. Multiple conceptual models in the model space (bottom) and predictions one can expect to obtain from alternative models (top). Mutually exclusive (non-overlapping models in the model space), collectively exhaustive (alternative models fill out the entire model space) (MECE) models are illustrated in a) which is the unobtainable goal of the model development process. b) mutually exclusive, non-collectively exhaustive models may underestimate uncertainty and lead to biased predictions. c) alternative models that are not mutually exclusive may lead to very similar predictions and thereby a false confidence in the result.

1.2.2 Model testing

In a model testing exercise the understanding that we have developed is challenged by comparing the model with new data not previously used in the model development. In hydrogeology this is sometimes referred to as “model validation”, although this term is avoided here as it implies the result of the exercise and that models can actually be validated (Oreskes et al., 1994). The objective of model testing is to gain confidence in the developed models. A nice side effect is that we may be able to reject some models, which is sometimes considered even more valuable (Beven, 2018; Hunt and Welter, 2010). Rejecting models reduces the uncertainty of predictions and the workload for future modelling exercises in the area as one less conceptual model will have to be considered, i.e. if model testing shows preferential recharge to an aquifer is not probable, we do not need to consider this in future modelling exercises.

Some concepts related to model testing are illustrated in Figure 1.3. Again, the grey box represents the plausible model space, while the coloured ones represent alternative models. In a successful model testing exercise, some models that were otherwise plausible, will be rejected because they are inconsistent with the testing data. From Figure 1.3a to Figure 1.3b two models are rejected thereby reducing the plausible model space and the predictive uncertainty of groundwater models based on the defined conceptual model.

In the exploration focused approach, we can afford to take an anti-realist viewpoint (Okasha, 2002) in that we believe many models can explain data (Figure 1.1). However, in a prediction focused approach, we are often pushed to take a realist viewpoint; that some of these models are better than others. We are pushed towards the realism viewpoint as the problem otherwise becomes too computational expensive (e.g. Mustafa 2020). What makes some models better than others is, however, still controversial.

In prediction focussed groundwater modelling, models are consequently sometimes removed from a model ensemble even if they are still considered plausible. Here the useful models are separated from models that are not useful. Whether models are useful often depends on the objective of the model workflow. The choice of models to be removed from the ensemble is based on an intercomparison of model performance, e.g. with a model selection technique or a Bayes Factor (Brunetti et al., 2017; Poeter and Anderson, 2005). In Chapter 3 we apply this approach and try to remove models from the ensemble by setting a threshold on the Bayes Factor.

In an exploration focussed approach however, models are only removed from the ensemble if they are inconsistent with data, i.e. the consistent models are separated from models that are inconsistent. The models are compared directly to data rather than to each other. We apply this approach in Chapter 4 where a Bayesian framework is combined with a falsification type approach. The falsification type approach consists of checking the models against independent data and rejecting models that are inconsistent.

Models that are not eliminated from the ensemble are confirmed conditional on the available data (Beven and Young, 2013). In an exploratory analysis, the more data the model is tested against the more certainty we can have that the model is likely going to be correct. In a prediction focussed approach the data used for testing might be restricted by the modelling objective.

Theoretically all models are wrong in that they are all very crude simplifications of reality, but the conditionally confirmed models are wrong within reasonable limits. They represent the models that are best suited to represent different aspects of the system.

Model-based hypothesis testing is limited by confirmation holism, that is we can only test collections of hypotheses and not individual hypotheses. If a model is found to be inconsistent with data, we will not know if the fault lies with the hypothesis we sought to test or somewhere else in the model. A hypothesis can therefore also only be conditionally rejected based on the assumptions made elsewhere in the model. In the model testing approaches in Chapter 3 and 4 we have focused on very simple models. As simple models have fewer assumptions, the assignment of the fault of a rejected model to the correct hypothesis is more likely than in a complex model with many assumptions.

In case all developed models are rejected (Figure 1.3c), a conceptual surprise has been uncovered. That is, none of the models in the ensemble are consistent with the data. A conceptual surprise should ideally lead to a complete overhaul of the conceptualization, rethinking the entire model structure. However, as this requires time and effort, sometimes the data that does not conform to our expectations is discounted unconsciously or the model is slightly changed through an ad-hoc modification to conform to the data. The latter approach can be taken because of confirmation holism, i.e. the fault is thought to lie somewhere else, not in the conceptual understanding. This is a convenient assumption as starting the modelling process over from the beginning is avoided.

In summary model testing approaches are important because they can increase the confidence in the models resulting from the model development exercise. Model testing in hydrogeology is, however, still relatively rare. In many of the studies reviewed in Chapter 2 alternative plausible models are removed from an ensemble without being falsified (realists' approach) and even fewer studies only remove models from the ensemble if they are falsified (anti-realists' approach) (Figure 1.1). This means developed models are rarely justified using independent data. While we still don't have a systematic approach to model development or discovery, we need to apply a systematic justification approach to make the groundwater modelling workflow a rational, logical process.

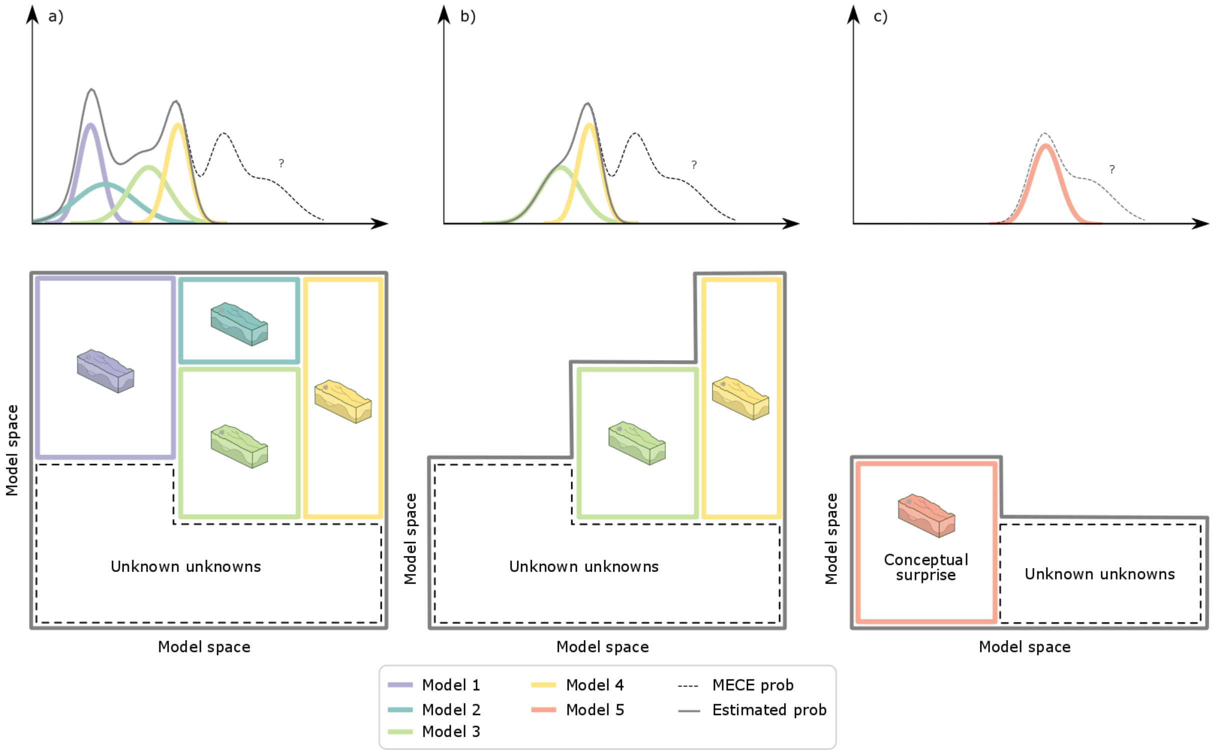

Figure 1.3. Influence of conceptual model testing on the plausible model space and predictions. a) mutually exclusive but not collectively exhaustive alternative models developed through a model development exercise. b) The result of a testing exercise where two models have been rejected from the initial ensemble in a), thereby reducing the plausible model space and range of predictions. c) All models from a) are rejected and a new model has therefore been defined; the conceptual surprise. Note that parts of the model space may still be unknown although a conceptual surprise has been uncovered.

1.3 FIELD SITE AND DATA COLLECTION

The methods presented in this thesis have been applied to Wildman River area, Northern Territory, Australia (Figure 1.4). Two major investigations in Wildman River area have been undertaken in recent years by the Department of Natural Resources in Northern Territory (Tickell and Zaar, 2017) and CSIRO as part of the Northern Australia Water Resource Assessment (NAWRA) (Turnadge et al., 2018a). The aim of both was to evaluate and improve understanding of the water resources as well as identifying areas suitable for groundwater dependent agricultural development. NAWRA further assessed the impact and risks of water resource and irrigation development by using a groundwater model.

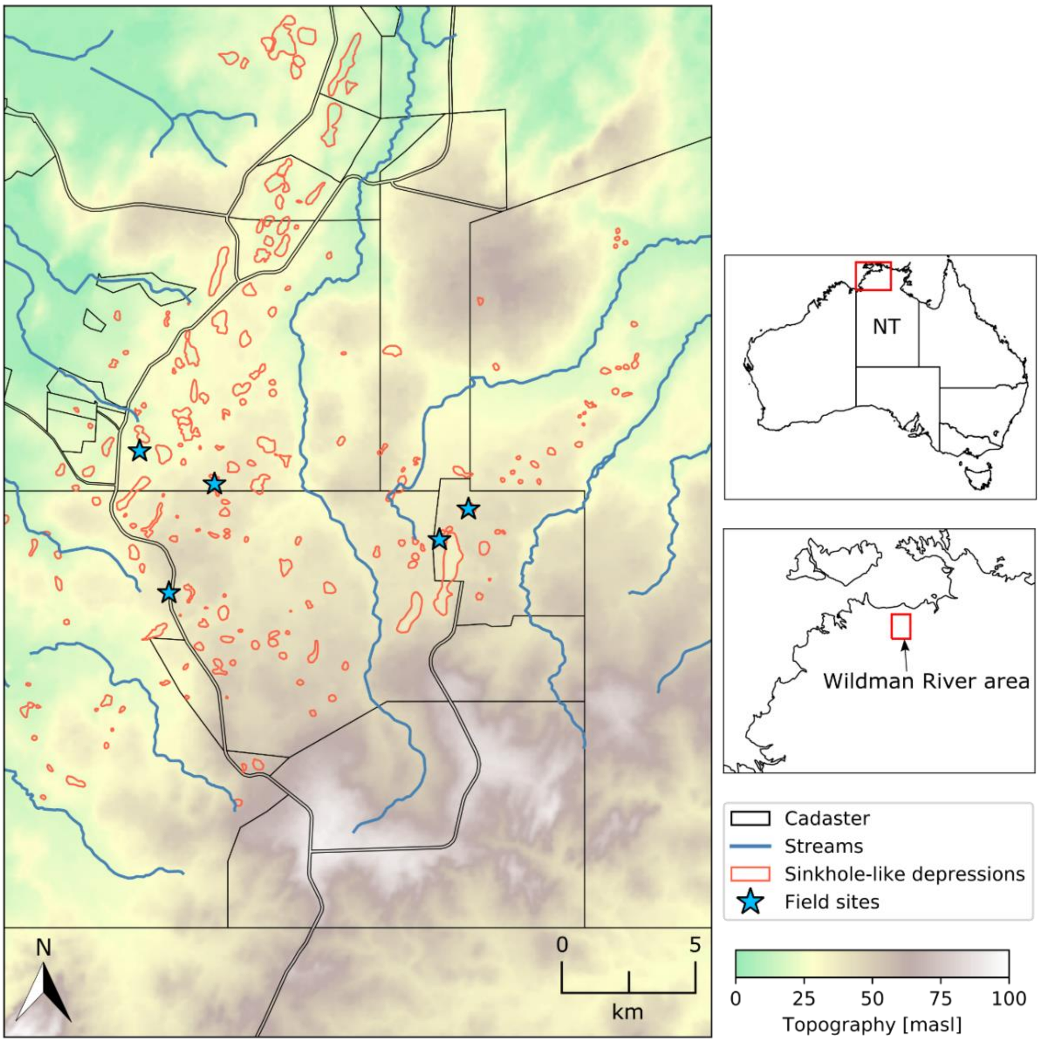

Figure 1.4. Overview of Wildman River area located in Northern Territory (NT), Australia. The field sites for the for Chapter 4 are marked with blue stars.

The Wildman River area was chosen as a case study for this thesis as the area still has potentially important hydrogeological features that may impact the susceptibility of shallow groundwater to developments at the surface. These features, i.e. the degree of surface water – groundwater interactions via streams and sinkhole-like depressions, have been the subject of previous research but several open questions remain regarding their hydrogeological conceptualisation. Some of these conceptual questions related to the water balance of the Wildman River area are addressed in Chapter 3, while the structure and functioning of sinkhole-like depressions are addressed in Chapter 4.

As part of Chapter 3, a field trip to Wildman River area (Figure 1.5) was conducted to collect water level measurements and to get a general hydrological and hydrogeological understanding of the area. A second field trip was conducted to collect essential data to investigate the sinkhole-like depressions (Chapter 4). Refraction seismic data, frequency domain electromagnetic induction (CMD data), sediment samples, water levels, topography and 360-degree photos were collected for five depressions, although not all the data was used in Chapter 4.

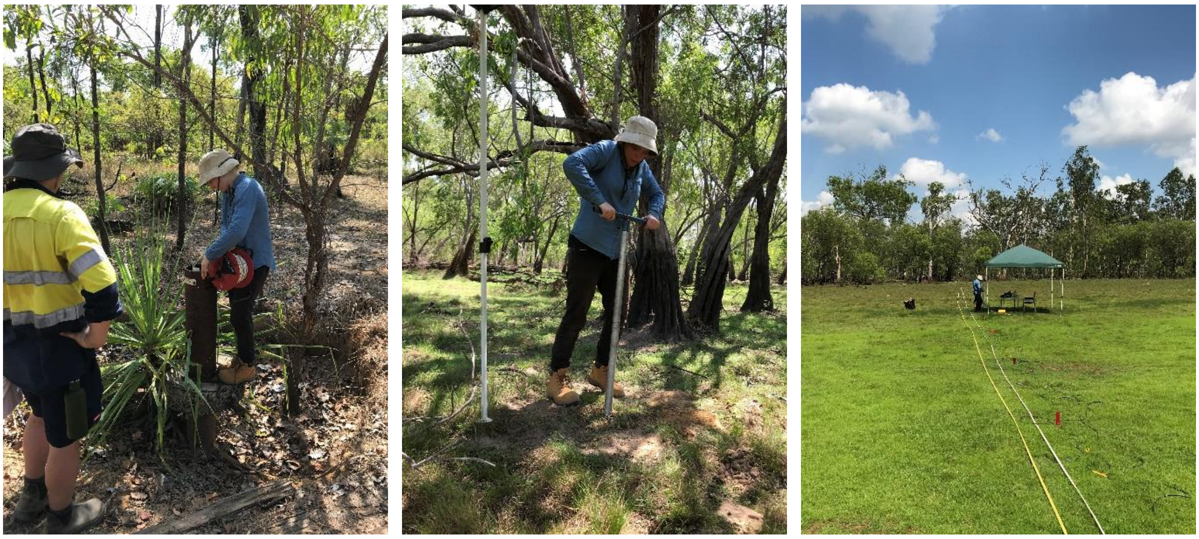

Figure 1.5. Field trips to Wildman River area. Collecting water level measurements (left), sediment samples (middle) and seismic refraction data (right).

1.4 RESEARCH AIM

Even if conceptual uncertainty is considered a major source of uncertainty in most groundwater modelling studies, no systematic approach exits to characterise and evaluate conceptual uncertainty. A systematic approach to evaluating conceptual uncertainty is necessary to increase the transparency and reproducibility of the groundwater modelling process and minimize the underestimation of uncertainty of model predictions. The aim of this thesis is to make conceptual model development and testing approaches in the multi-model approach more accessible by developing workflows that collectively provide a comprehensive, objective, and repeatable workflow to develop and test conceptual models.

While the focus of the application of the conceptual model uncertainty workflow is on a hydrogeological system, conceptual uncertainty is not specific to hydrogeology. Insights from this thesis can be transferred to other model-based disciplines where assumptions on the modelled system functioning are necessary like economy, biology and meteorology.

The specific aims of the thesis are to:

1. Identify current approaches, unify scattered insights and improve the methodology of hydrogeological conceptual model development for characterisation of conceptual uncertainty.
2. Identify current approaches, unify scattered insights and improve the methodology of hydrogeological conceptual model testing to increase confidence in conceptual models and subsequent groundwater flow model outputs.
3. Increase hydrogeologic system understanding of key features in Wildman River area such as the water balance and the recharge process by applying approaches 1 and 2.

The way the aims of this thesis maps to the thesis structure is presented in Table 1.1.

1.5 STRUCTURE OF THIS THESIS

Chapter 2 provides an overview of how conceptual uncertainty have been considered and evaluated in international literature. The chapter goes beyond just reviewing the literature by defining for the hydrogeological community an essential future pathway for simulating conceptual uncertainty. It focuses on the conceptual uncertainty in hydrogeological models, the presented concepts are however generally applicable to all spatio-temporal dynamical environmental systems models. This chapter has been published in the peer-reviewed Journal of Hydrology (Enemark et al., 2019a).

Chapter 3 presents an approach to model-based Bayesian hypothesis testing in a simple additive stochastic groundwater balance model, which involves optimization of a model in function of both parameter values and conceptual model through trans-dimensional sampling. The method was demonstrated on a water balance model for the Wildman River area. Although none of the conceptual models could be rejected, more confidence was gained in the water balance predictions. This chapter has been published in the peer-reviewed journal Water (Enemark et al., 2019b).

Chapter 4 proposes an approach to systematic hydrogeological conceptual model development and testing. The method is applied to the Wildman River area where sinkhole-like depressions are tested using remote sensing and geophysical data to evaluate whether they can act as conduits for recharge. Despite focussing on a very specific conceptually uncertain component in Wildman River area, the presented methodology on systematically testing conceptual models is generally applicable. Chapter 5 is submitted to the peer-reviewed journal Water Resources Research.

Chapter 5 summarises and infer conclusions from the thesis chapters and provides an outlook for applications and future investigations.

The appendices A-C provide chapter specific additional information for chapters 3-5, while Appendix D provides a publication list resulting from the PhD project.

Table 1.1. Overview of thesis structure (Section 1.4) in relation to the research aims (Section 1.4).

	Research aim 1	Research aim 2	Research aim 3
Chapter 2	Identify current approaches, unify scattered insights, identify current challenges.		
Chapter 3	Mutually exclusive models in factorial approach based on study site specific literature review.	Realism approach. Making testing more accessible by applying simple water balance model.	Improve confidence in water balance.
Chapter 4	Mutually exclusive models in factorial approach based on general literature review and drawing analogies to the study site.	Anti-realism approach. Making testing more accessible by presenting a framework to model testing using an anti-realism approach.	Improve conceptualization of sinkhole-like depressions.

Chapter 2: Hydrogeological Conceptual Model Building and Testing: A Review

Trine Enemark, Luk JM Peeters, Dirk Mallants, Okke Batelaan

Published in *Journal of Hydrology*[1]

[1] Enemark, T., Peeters, L. J. M., Mallants, D., & Batelaan, O. (2019). Hydrogeological conceptual model building and testing: A review. Journal of Hydrology, 569. https://doi.org/10.1016/j.jhydrol.2018.12.007

2.1 ABSTRACT

Hydrogeological conceptual models are collections of hypotheses describing the understanding of groundwater systems and they are considered one of the major sources of uncertainty in groundwater flow and transport modelling. A common method for characterizing the conceptual uncertainty is the multi-model approach, where alternative plausible conceptual models are developed and evaluated. This review aims to give an overview of how multiple alternative models have been developed, tested and used for predictions in the multi-model approach in international literature and to identify the remaining challenges.

The review shows that only a few guidelines for developing the multiple conceptual models exist, and these are rarely followed. The challenge of generating a mutually exclusive and collectively exhaustive range of plausible models is yet to be solved. Regarding conceptual model testing, the reviewed studies show that a challenge remains in finding data that is both suitable to discriminate between conceptual models and relevant to the model objective.

We argue that there is a need for a systematic approach to conceptual model building where all aspects of conceptualization relevant to the study objective are covered. For each conceptual issue identified, alternative models representing hypotheses that are mutually exclusive should be defined. Using a systematic, hypothesis based approach increases the transparency in the modelling workflow and therefore the confidence in the final model predictions, while also anticipating conceptual surprises. While the focus of this review is on hydrogeological applications, the concepts and challenges concerning model building and testing are applicable to spatio-temporal dynamical environmental systems models in general.

2.2 INTRODUCTION

Groundwater model conceptualization is a crucial first step in groundwater model development (Anderson et al., 2015a). It provides a systematic, internally consistent overview of system boundaries, properties and processes relevant to the research question, bridging the gap between hydrogeological characterization and groundwater modelling.

As the conceptualization is related to the fundamentals of the problem definition, it is considered one of the major sources of uncertainty in numerical groundwater modelling (Gupta et al., 2012). Estimating parameters through calibration with an inadequate conceptual model may lead to biased parameter values (Doherty and Welter, 2010). Biased parameter values are especially problematic when extrapolating to predictions that are of a different type than the calibration data, represent a different stress regime, or have a longer timeframe than the calibration period (White et al., 2014). Not accounting for conceptual model uncertainty can potentially greatly underestimate total uncertainty and give false confidence in model results, as vividly illustrated in Bredehoeft (2005).

To develop conceptual models, two major approaches have been traditionally applied: (i) the consensus model approach (Brassington and Younger, 2010) and (ii) the multi-model approach (Neuman and Wierenga, 2003) (Figure 2.1). The development of conceptual models is based on the available geological and hydrological information, which are observed data, such as water levels, borehole information and tracer concentrations, but often also include a component of soft knowledge, such as geological insights or expert interpretation.

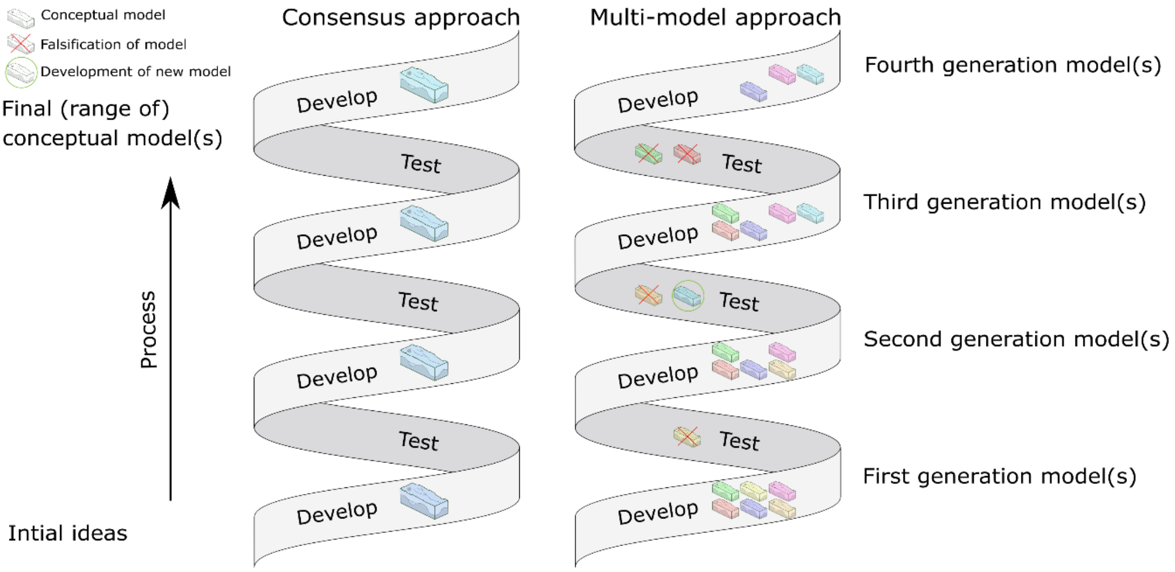
Conceptual model
Falsification of model
Development of new model
Final (range of) conceptual model(s)
Process
Intial ideas
Consensus approach
Multi-model approach
Develop
Test
Fourth generation model(s)
Third generation model(s)
Second generation model(s)
First generation model(s)

Figure 2.1. Iterative process for the conceptual modelling process via the consensus or multi-model approach. Modified from Environment Agency (2002) and Suzuki et al. (2008). Each model test step involves introducing new data and thereby identifying new plausible models uncovering conceptual surprises, and rejecting other models that are inconsistent with the new data.

In the single consensus conceptual model approach all available observations and knowledge is iteratively integrated into a single conceptual model (Barnett et al., 2012; Izady et al., 2014), providing a staircase of confidence (Gedeon et al., 2013). In this case, the conceptual model represents the current consensus on system behaviour (Brassington and Younger, 2010).

As illustrated in Schwartz et al. (2017), conceptual model uncertainty is generally accounted for in the consensus approach by increasing the complexity of the model. Increasing complexity effectively turns conceptual model uncertainty into parameter uncertainty by adding more processes to the model and/or increasing resolution in space and time. Increasing the degrees of freedom means that non-uniqueness increases, which is often balanced through optimal model complexity favouring the simplest model that can adequately reproduce historical conditions (Young et al., 1996). The main advantage is that it comprehensively captures conceptual issues in the model. The main drawback is that models quickly become intractable and too computationally demanding to carry out parameter inference. Another mechanism that is often applied to account for conceptual uncertainty, is conservatism, favouring the conceptualization that will result in the largest impact (Wingefors et al., 1999). Although inherently biased, the main advantage is that introducing conservative assumptions make the problem tractable and provides confidence that the simulated impacts are not underestimated. The largest drawback however, is that conservative assumptions depend on the type of impact investigated, may not be internally consistent and can lead to missed opportunities (Freedman et al., 2017).

The alternative to the consensus approach is the multi-model approach, in which an ensemble of different conceptualizations is considered throughout the model process in parallel rather than sequentially. This approach reflects that the hydrogeological functioning of an aquifer system can be interpreted in different ways, especially if the available data is scarce (Anderson et al., 2015a; Beven, 2002; Neuman and Wierenga, 2003; Refsgaard et al., 2006). In the multi-model approach the aim is not to find the single best model, but to find an ensemble of alternative conceptual models, each with a different hypothesis on system behaviour. As depicted in Figure 2.1, this is also an iterative process, in which conceptual models are removed from the ensemble when they are falsified by increased knowledge or data, and where conceptual models are added when new data or insights prompt the development of a new hypothesis on model behaviour.

In the consensus approach, once committed to a particular conceptualization, there is considerable inertia to change it as this would often involve a complete overhaul of the numerical model (Ferré, 2017). However, in the multi-model approach, given alternative conceptual models are developed and evaluated in parallel, it aids in solving the problem of conceptual "surprises" (Bredehoeft, 2005) as they are sought out. Even though the multi-model approach is less prone to conceptual surprises than the consensus approach, it is not exempt from it. Using statistical terminology, as explained by Neuman (2003), both the consensus approach and the multi-model approach are prone to Type I errors (underestimating model uncertainty by undersampling the model space) and Type II errors (relying on invalid model(s)). However, by using the multi-model approach we are less likely to commit either.

This paper aims to provide an overview of the current status of the international literature on using multiple conceptual models in groundwater modelling. Reviews of the multi-model approach to date, such as Diks and Vrugt (2010), Schöniger et al. (2014), and Singh et al. (2010) mainly focus on the evaluation of multiple models and summarising of model results. Much less attention has been devoted to approaches that systematically develop and test different conceptual models. This review is therefore organized around the following four research questions:

1. What is conceptual model uncertainty?
2. How are alternative conceptualizations developed?
3. How can alternative conceptualizations be tested?
4. How are different conceptualizations used for predictions?

Each section provides an overview of approaches in published studies, summarized in Table A.1 and Table A.2, and remaining challenges. While this review will focus on applications in a hydrogeological context, the concepts and challenges concerning model building and testing are applicable to spatio-temporal dynamical environmental systems models in general.

2.3 WHAT IS CONCEPTUAL MODEL UNCERTAINTY?

Anderson and Woessner (1992) and Meyer and Gee (1999) define a conceptual model as a pictorial, qualitative description of the groundwater system in terms of its hydrogeological units, system boundaries (including time-varying inputs and outputs), and hydraulic as well as transport properties (including their spatial variability). The conceptual model is often seen as a hypothesis or a combination of hypotheses for the aspects of the groundwater system that are relevant to the model objective.

Table A.1 provides a review of internationally peer reviewed publications that explicitly consider hydrogeological conceptual model uncertainty. These 59 studies have been identified from the Google Scholar database, where the search term “groundwater model” is combined with “conceptual model uncertainty”, “structural model uncertainty”, “alternative conceptual models” or “multi-model approach”. Only studies that include alternative conceptual models developed for groundwater modelling, for the purpose of either increasing system understanding or characterizing conceptual uncertainty, have been included. This list is considered to be representative of the treatment of conceptual model uncertainty through the multi-model approach in groundwater research in the last two decades. It is beyond the scope of this review to address the consensus conceptual model building approach. For each study, Table A.1 provides a short summary of the alternative conceptualizations, whether or not the objectives are explicitly defined and which aspects of the conceptualization are considered.

In this section we discuss what is included in model conceptualization, how this needs to be linked to the objective of the modelling and the linguistic ambiguity in discussing conceptual model uncertainty.

2.3.1 Conceptual model aspects

Gupta et al. (2012) outlines five formal stages in the model building process: i) Conceptual Physical Structure, ii) Conceptual Process Structure, iii) Spatial Variability Structure, iv) Equation Structure and v) Computational Structure. The first two steps are part of the conceptual model, the third and fourth are part of the mathematical model and the last step is the computational model. This review will focus on the first two steps, as well as the Spatial

Variability Structure (Figure 2.2). The latter is included in our discussion of aspects of conceptualization as some studies in Table A.1 consider alternative models of the Spatial Variability Structure as conceptual uncertainty.

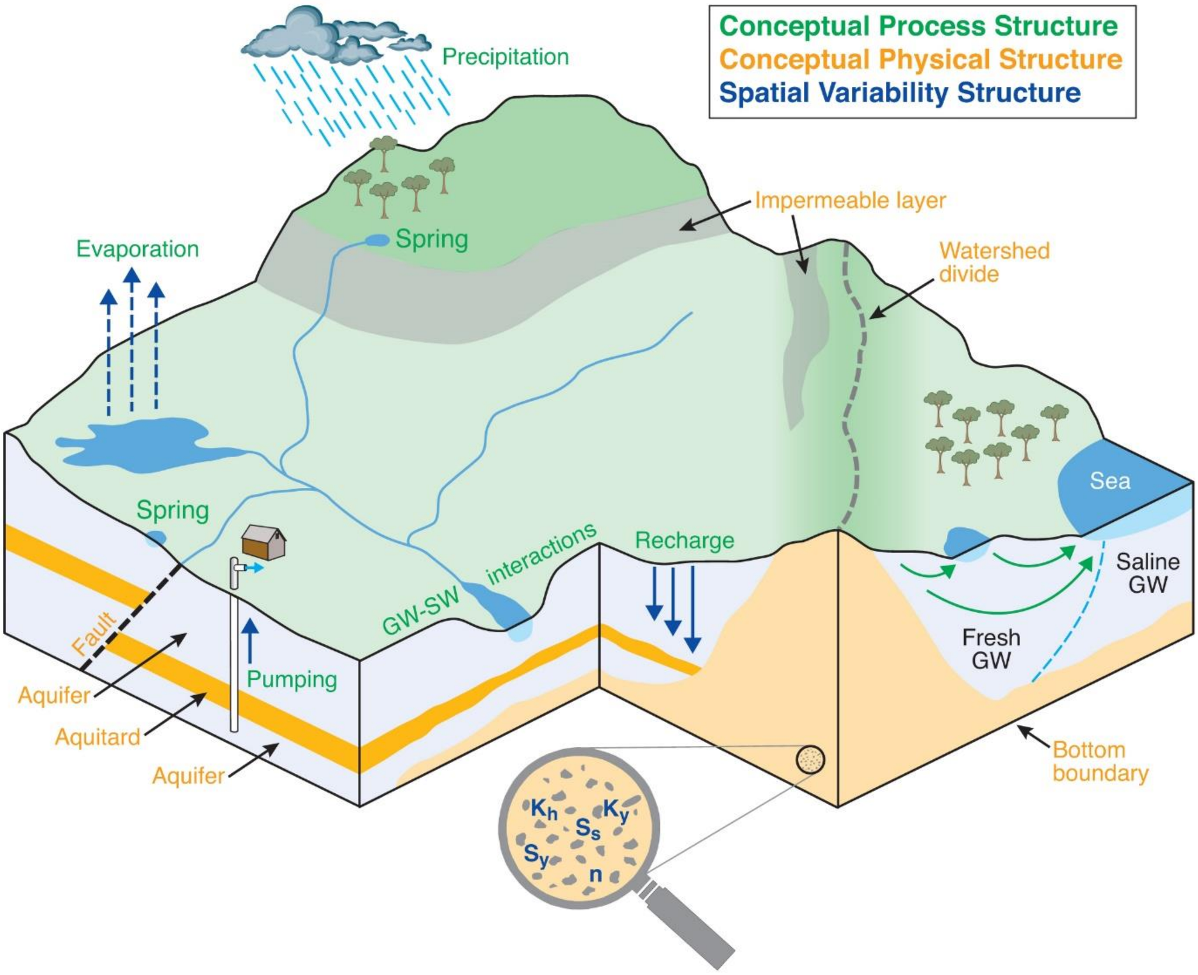

Figure 2.2. Elements of a conceptual model. Items in green illustrate the Conceptual Process Structure, while items in blue illustrate the Spatial Variability Structure represented in the magnifying glass (K_h = horizontal hydraulic conductivity, K_v = vertical hydraulic conductivity, n=porosity, S_s = Specific storage, S_y = Specific yield). Items in orange illustrate the Conceptual Physical Structure represented the system geometry and hydrostratigraphy.

The Conceptual Physical Structure captures the hydrostratigraphy as well as the horizontal and vertical extent of the system (respectively a watershed divide and an impermeable bottom boundary in Figure 2.2). The Conceptual Physical Structure further defines the hydrostratigraphic units and their extent, the barriers and/or conduits to groundwater flow (faults) and the compartmentalisation of the groundwater system into aquifers and aquitards. The Spatial Variability Structure is the description of the time-invariant hydraulic properties of the system and their spatial variability (magnifying glass in Figure 2.2). The Conceptual Process Structure contains the boundary conditions that are time variant, such as heads and fluxes in and out of the system. These can be externally controlled and largely independent from the groundwater system dynamics (e.g., rainfall, pumping rates, drainage levels for mine dewatering, lateral zero-flow boundary) or internally controlled and largely dependent on the groundwater system dynamics (e.g., surface water-groundwater interaction, evapotranspiration).

2.3.2 Modelling objective

Despite being identified as the crucial first step in any modelling study (Anderson et al., 2015a; Barnett et al., 2012; Brassington and Younger, 2010), only 33 out of 59 studies explicitly define the purpose or objective of the model in the introduction of the paper. This is especially relevant as some conceptualization aspects (such as detailed description of spatial variability of hydraulic properties) might be important to one type of prediction (e.g., travel time distribution), but might be less relevant to another type of prediction (e.g., hydraulic head distribution) (Refsgaard et al., 2012; Zhou and Herath, 2016). Alternative conceptualizations are for instance directly linked to model objectives when multiple conceptual models are developed to increase system understanding (Passadore et al., 2011) or aid in water management strategy (Højberg and Refsgaard, 2005). Many of the studies in which a model objective is not explicitly defined, are focused on method development, such as combining model averaging techniques (Rojas et al., 2008), comparing ranking strategies (Foglia et al., 2007) or model selection (Poeter and Anderson, 2005).

2.3.3 Linguistic uncertainty

There is considerable linguistic ambiguity in describing the uncertainty of groundwater system conceptualization. A prime example is the term ‘structural uncertainty’, which can indicate uncertainty in geological structure, as in Refsgaard et al. (2012), or can indicate the number and type of processes represented in the numerical model, as exemplified in Clark et al. (2008).

Furthermore, as argued in Nearing et al. (2016) any adequate model should encode all uncertainties to consider, i.e. the known unknowns. The name ‘multi-model approach’ is therefore somewhat misleading. The multiple models in the multi-model approach are samples of the overall plausible model choices that should characterize the conceptual uncertainty. This is no different than sampling parameters over a feasible range to characterize the parameter uncertainty. In this definition, the multiple models in the multi-model approach therefore only represent a single model characterizing known unknowns.

The linguistic uncertainty has led to a wide variation in what is considered to be conceptual model uncertainty (Table A.1). This varies from changing the hydraulic conductivity zonation extent and number (Carrera and Neuman, 1986; Foglia et al., 2007; Lee et al., 1992; Meyer et al., 2007; Poeter and Anderson, 2005) to considering different process representations (Altman et al., 1996; Aphale and Tonjes, 2017). Classifications of sources of uncertainty, such as presented in Walker et al. (2003), Refsgaard et al. (2006) or Vrugt (2016), often distinguish between model structure uncertainty (incomplete understanding and simplified description of modelled processes), parameter uncertainty (parameter values) and input uncertainty including scenario uncertainty (external driving forces). In groundwater model conceptualization, the distinction between these classes is not well defined. For example, should changing the Spatial Variability Structure of hydraulic conductivity, such as in Castro and Goblet (2003), Rogiers et al. (2014), or Linde et al. (2015), be considered conceptual or parameter uncertainty?

Suzuki et al. (2008) provides a more pragmatic classification in which differentiation is made between first-order uncertainties (conceptual) and lower-order uncertainties. Lower-order uncertainties are aleatory and can be modelled stochastically, while conceptual uncertainties are epistemic and are characterized by alternative models. Common in both the consensus model approach and the multi-model approach is that lower-order uncertainties are modelled stochastically within each conceptualization. For example, Hermans et al. (2015) uses different training images to describe spatial variability of hydraulic conductivity with multiple-point geostatistics; this can be considered a first-order uncertainty. The lower-order uncertainty is then the stochastic realisations of each training image. Likewise, changing the boundary from a no-flow to a head dependent boundary in Mechal et al. (2016) is first-order uncertainty, while changing the value of the head-dependent boundary in Aphale and Tonjes (2017) is considered a characterization of lower-order uncertainty.

2.3.4 Summary of what is considered conceptual model uncertainty

Groundwater system conceptualization is a collection of hypotheses describing the understanding of the different aspects of the groundwater system that are important to the modelling objective. Conceptual model uncertainty is the uncertainty due to the limited data and knowledge about a groundwater system. It is the first-order, epistemic uncertainty that is generally considered reducible but cannot be characterized by continuously varying a variable. Linguistic ambiguity and vague definitions of what constitutes conceptual uncertainty however hinders transparent discussions of this major source of uncertainty. We will therefore adopt the terminology of Suzuki et al. (2008) and focus on first-order uncertainty.

2.4 HOW ARE DIFFERENT CONCEPTUALIZATIONS DEVELOPED?

Not only is there a wide variety of conceptual model aspects, there is also a wide variety of ways to generate different conceptualizations (Table A.1). Generating different conceptualizations has not received much attention in the literature and guidance is likewise limited. Neuman and Wierenga (2003) discuss different approaches in developing alternative conceptualization and suggest building alternative models until no other plausible explanations can be identified. Similar to this approach, Refsgaard et al. (2012) introduced the concept of the Mutually Exclusive and Collectively Exhaustive (MECE) criterion to hydrogeology. In order to be mutually exclusive, conceptual models have to be completely disjoint and represent independent hypotheses about the groundwater system. In order to be collectively exhaustive, the entire range of plausible conceptual models needs to be defined, including the unknown unknown plausible models. The unknown unknowns are the conceptual models that current data has not yet uncovered and will lead to conceptual surprises if they are. It has been acknowledged by several authors that defining a collectively exhaustive range is impossible in practice (e.g. Ferre, 2017; Hunt and Welter, 2010; Refsgaard et al., 2012).

While the concepts and advice in Neuman and Wierenga (2003) and Refsgaard et al. (2012) are sound and highly relevant, few of the studies in Table A.1 adhere to them. From the studies of Table A.1, three main strategies are identified in developing alternative conceptualizations; (i) Varying Complexity, (ii) Alternative Interpretations and (iii) Hypothesis Testing. These strategies are illustrated in Figure 2.3.

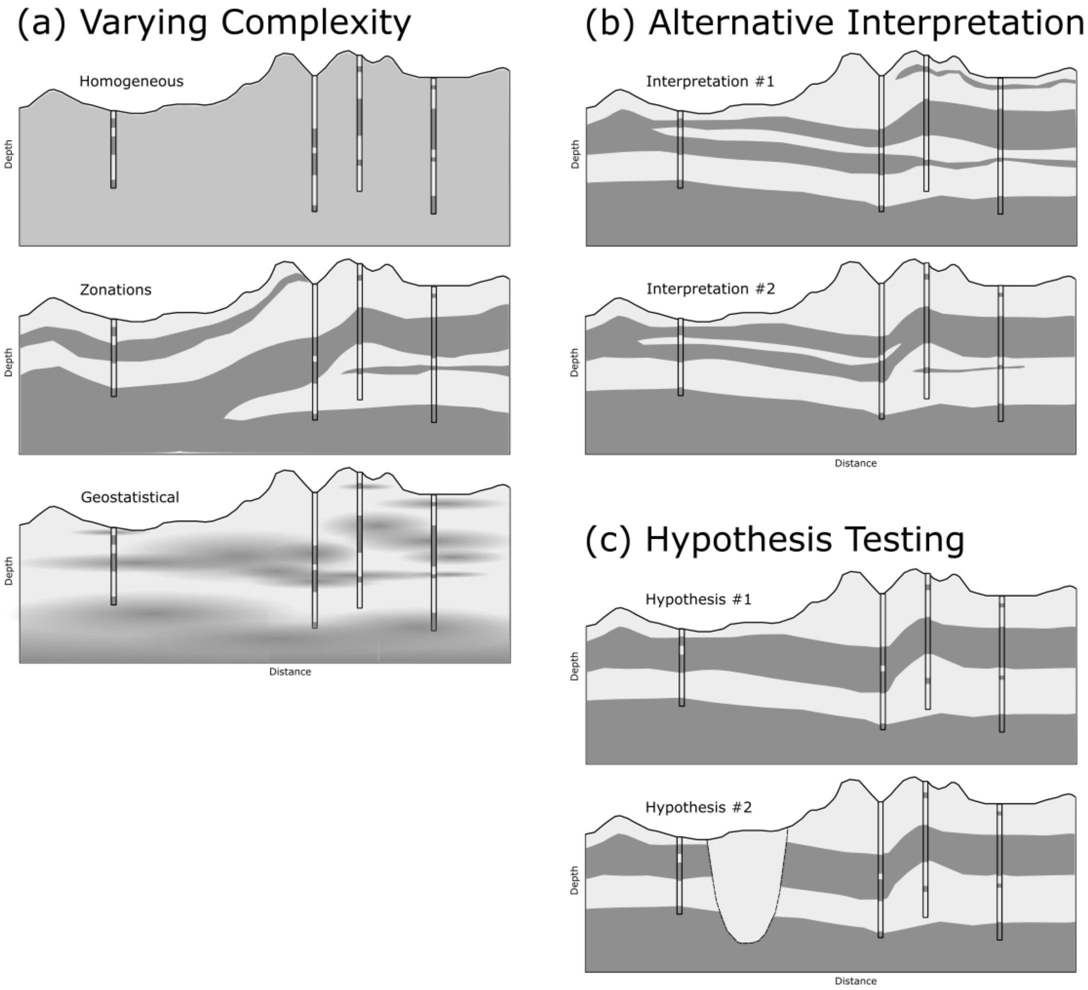

Figure 2.3. Conceptual model development approaches in the multi-model approach. Illustration of how different conceptualizations of the Conceptual Physical Structure could take shape if based on the same data (boreholes in this case) through Varying Complexity (a), Alternative Interpretation (b) or Hypothesis Testing (c) strategy. Based on illustrations of alternative models in Harrar et al. (2003), Schöniger et al. (2015), Seifert et al. (2008) and Troldborg et al. (2007).

In the Varying Complexity strategy, alternative models are generated by gradually increasing or decreasing the complexity of the same base conceptualization. In Figure 2.3 this is illustrated by describing the hydraulic property variability in an aquifer system either as (i) homogeneous units, (ii) zonation or (iii) a spatially continuous parameterization. The adequate complexity is typically evaluated based on the modelling goal (Höge et al., 2018; Zeng et al., 2015), the available data (Schöniger et al., 2015a), or the informative model complexity (Freedman et al., 2017). The underlying base conceptualization is not questioned and it is, often implicitly, assumed that all conflict between observed and simulated data is due to the inability to capture the full complexity of the groundwater system in the numerical model. The Varying Complexity strategy does not fit well in the MECE paradigm as different levels of complexity in implementing the same conceptualization do not ensure mutually exclusive hypotheses.

The Alternative Interpretation strategy consists of generating an ensemble of conceptualizations by different interpretations. Figure 2.3 illustrates this as two different hydrostratigraphic interpretations of the same borehole data set, independent by being interpreted by different teams who have no knowledge about the each other’s interpretation (e.g. Harrar et al., 2003; Hills and Wierenga, 1994). Compared to the Varying Complexity strategy, the Alternative Interpretation strategy has the advantage that the ensemble can include very different base conceptualizations (e.g. Refsgaard et al., 2006). However, the conceptualizations may end up being very similar and it is difficult to ensure that independent interpretations are mutually exclusive.

In the Hypothesis Testing strategy, as advocated by Beven (2018), an ensemble of models is generated by stating different hypotheses about the system. Rather than multiple teams formulating their best interpretation of the same data in the Alternative Interpretation strategy, the Hypothesis Testing strategy involves the same team aiming to maximise the difference between alternative conceptualizations, while still adhering to the same dataset. In Figure 2.3 this is exemplified through the presence or absence of a palaeovalley in two alternative conceptualizations. Both alternatives are consistent with the borehole data, but the interpretation with the palaeovalley present may be considered less likely. The chances are slim that such a vastly different conceptualization would be part of an ensemble generated through the Alternative Interpretation strategy, where only the most likely model is sought. None of the three strategies guarantees that the ensemble of models developed is collectively exhaustive, but it is more likely for Hypothesis Testing to generate an ensemble of mutually exclusive models.

The next sections review model building approaches and are structured around the three key components of the conceptual model illustrated in Figure 2.2; Conceptual Physical Structure (Section 2.4.1), Spatial Variability Structure (Section 2.4.2), and Conceptual Process Structure (Section 2.4.3). The focus is on different approaches to building multiple conceptual models within these three aspects and how the different strategies to multi-model building have been applied (Figure 2.3). Finally, Section 2.4.4 discusses assigning prior probabilities to alternative models.

2.4.1 Conceptual Physical Structure

Table A.1 lists several examples where the Conceptual Physical Structure of conceptual models has been tested through the Alternative Interpretation and the Hypothesis Testing strategy. Using an Alternative Interpretation strategy approach, five alternative hydrostratigraphic models were generated by five different (hydro)geologists in the study by Seifert et al. (2012) resulting in different number of layers, proportions of sand and clay in the quaternary sequence and the location of a limestone surface. Using the Hypothesis Testing strategy, Troldborg et al. (2007) developed three different models by assuming different depositional histories and thereby different number of layers in the models.

While it is possible to test a global geometrical hypothesis about the Conceptual Physical Structure (e.g. Troldborg et al. (2007)), it is more common to test specific geometrical features through local hypotheses. A local hypothesis can for instance test the presence of a palaeovalley (Seifert et al., 2008), the connection between two aquifers (La Vigna et al., 2014), or the extent of an aquifer (Aphale and Tonjes 2017). If one of the hypotheses is falsified in these studies, the system understanding will improve in regard to that specific feature.

2.4.2 Spatial Variability Structure

Spatial Variability Structure is the component of the conceptual model that is most often included in a multi-model approach. Because hydraulic and transport properties are often scale-dependent and the adequate level of complexity depends on the modelling purpose, the description of properties is often tested by developing models with the Varying Complexity strategy. The strategy is applied either through dividing the study area into different zones of homogeneous hydraulic conductivities, so alternative representations can be generated by combining the different zones (e.g. Foglia et al., 2007), or by representing the geology in different conceptual models as homogenous, layered/zoned, or as heterogeneous (e.g. Schöniger et al., 2015).

In the INTRAVAL Las Cruces trench experiment five different modelling teams developed unsaturated zone flow and transport models using the Alternative Interpretation strategy (Hills and Wierenga, 1994). Despite differences between the models, such as isotropic/anisotropic and spatially uniform/heterogeneous soil properties, none of the models was clearly superior considering several performance criteria.

Geostatistical variogram based approaches facilitate the stochastic generation of many pixel-based *K* realizations based on the same data and assumptions to characterize the lower-order uncertainty. Hypothesis Testing strategy has been applied assuming different variogram models to represent the *K* variation within the system (Samper and Neuman, 1989; Ye et al., 2004). Rather than defining different facies variogram, Pham and Tsai (2015; 2016) used three different variogram based geostatistical approaches (indicator kriging, indicator zonation and general parameterization (Elshall et al., 2013)) to describe the variation between clay and sand units as smooth or sharp.

In the multipoint geostatistics approach (MPS) (Strebelle, 2002) different conceptualizations can be represented by adopting different training images using the Hypothesis Testing strategy. Studies that have applied the MPS approach using more than one training image in groundwater modelling are still rare but include studies by He et al. (2014), Hermans et al. (2015) and Linde et al. (2015).

Groundwater flow through fractured rock aquifers complicates the conceptualization as the groundwater flow occurs through both matrix and fractures. Selroos et al. (2002) considered e.g. stochastic continuum models and discrete fracture networks as alternative conceptualizations of fractured rock in Sweden; the models were shown to have different results in terms of solute transport behaviour

2.4.3 Conceptual Process Structure

The Conceptual Process Structure is the component in the conceptual model that is considered least in the multi-model approaches in the analysed studies (Table A.1). According to Gupta et al. (2012) this lack of attention in literature is mainly due to the process description typically being assumed to be complete. However, as illustrated by examples in Bredehoeft (2005), conceptual surprises might also occur for the Conceptual Process Structure as well as for the other components of the conceptual model.

Among the many boundary conditions imposed on a groundwater model, groundwater recharge is by far the one that has received most attention in the literature. A number of methods exist for calculating groundwater recharge that take into account different sources of information (Doble and Crosbie, 2017; Scanlon et al., 2002) which can lead to different estimates of recharge when used in an Alternative Interpretation strategy approach. Ye et al.

(2010) used the Maxey-Eakin method, the chloride mass balance method and the net infiltration method to derive different estimates of recharge to assess the conceptual uncertainty. Each of the different interpretation methods resulted in a different spatial distribution of recharge.

Different levels of model complexity have often been used across different spatial scales, such as for groundwater recharge estimation (Doble and Crosbie, 2017). Models range from simplified heuristic models at a global scale (Döll and Fiedler, 2008), simple 1-D bucket models for regional scale areas (Flint et al., 2000) to more complex numerical solutions of Richards' equation at the field scale (Leterme et al., 2012; Neto et al., 2016). Nettasana (2012) tested the complexity of zonation of recharge by defining recharge based only on soil type in one model and in another model both on soil type and land use.

The Hypothesis Testing approach for recharge estimation mainly focuses on a specific feature (Kikuchi et al., 2015; Rojas et al., 2010a). Aphale and Tonjes (2017) investigate the effect of a landfill on local recharge with three different hypotheses. Hypothesis Testing for lateral boundary conditions has been applied to lateral exchange flux with adjacent aquifers (Lukjan et al., 2016; Mechal et al., 2016; Nettasana, 2012). Kikuchi et al. (2015) test the existence of underflow through a subsurface zone into an adjacent basin.

2.4.4 Assigning a prior probability

A crucial aspect in any Bayesian modelling approach is assigning the prior probabilities. This prior is based on an initial understanding of the probability of a model related to the alternative models and is updated when additional data is introduced in the model testing step (Section 2.5). The assigned prior for the reviewed studies are presented in the first column of Table A.2.

In order to be objective and unbiased, different conceptual models are often considered to be equally likely, uninformed by data or knowledge. From the 26 studies in Table A.2 that assign a prior probability, 21 use a uniform, and thus uninformed, prior probability. Prior probabilities do however have a large influence on the posterior probability if the data used for updating the prior has limited information content. Rojas et al. (2009) showed that including proper prior knowledge about the conceptualizations increased predictive performance when compared to assigning uninformed priors. Additionally, uninformed priors are not consistent with the Hypothesis Testing approach, as shown in Figure 2.3c. If no other palaeovalleys were observed in the area, the palaeovalley hypothesis would be possible, but unlikely. A uniform prior probability would assign each hypothesis equal likelihood, which would not be appropriate.

In the reviewed studies the prior has been based on expert opinion, data consistency and model complexity. For instance, using expert opinion in the study by Ye et al. (2008) the prior probability was based on expert’s belief in alternative recharge models considering the consistency with available data and knowledge. Systematic expert elicitation is a well-established technique in environmental risk assessment and modelling (Krueger et al., 2012) to formalize expert belief into model priors. There are however few published studies on expert elicitation in groundwater conceptualization context. Elshall and Tsai (2014) used data consistency to inform the prior probability by basing it on calibration of hydrofacies using lithological data. Finally, using model complexity to inform the prior, in the study by Ye et al. (2005) higher probabilities were assigned to favour models with fewer parameters. This was also suggested by Rojas et al. (2010a) as a means of penalizing increased complexity. Nearing et al. (2016) argues that assignment of probabilities should not be based on a single component of the model but rather be based on the whole model. In the reviewed literature the priors have however, only been based on individual components.

2.4.5 Remaining challenges

The review of studies in Table A.1 has shown that alternative models have been developed either by i) varying complexity of model description, ii) making alternative interpretations or iii) stating different hypotheses about the groundwater system. The goal of the multi-model development process is to define a mutually exclusive, collectively exhaustive range of models in which the true unknown model exists and where the risk of uncovering a conceptual surprise is zero. This is obviously unattainable, and we therefore discuss the remaining challenges next.

First, Table A.1 shows that studies typically focus on exploring different hypotheses for a single aspect of the model (Conceptual Physical/Conceptual Process/Spatial Variability Structure). Only 5 out of 59 papers consider all three aspects simultaneously (Aphale and Tonjes, 2017; Foglia et al., 2013; Mechal et al., 2016; Rojas et al., 2010a; Ye et al., 2010). For the range of models to be collectively exhaustive, all conceptually uncertain aspects must be considered.

Second, the study objective is not always considered when alternative models are developed for the multi-model approach (Table A.1). Models should encapsulate the behaviour that is important to the modelling objective (Jakeman et al., 2006), and The same should be true when characterizing conceptual uncertainty. On the other hand, “what may seem like inconsequential choices in model construction, may be important to predictions” (Foglia et al., 2013). To avoid ignoring the inconsequential model choices, the model objective should be used to guide the development of alternative models. This does imply that ensembles are not necessarily the same for all model objectives (Haitjema, 1995).

Third, alternative conceptual models are not always defined as mutually exclusive (i.e. if model A is true, models B and C are false). Falsification, which is welcomed in the multi-model approach (Beven, 2018), will increase system understanding (Beven and Young, 2013), but how much will depend on how the conceptual models are defined. In the Alternative Interpretation and Varying Complexity strategy, the models are not necessarily mutually exclusive in the sense that they do not represent different ideas about the groundwater system. In the Varying Complexity approach, alternative models are generated based on the same conceptual model represented in different complexities. A risk in the Alternative Interpretation strategy is that alternative models are almost identical in terms of understanding of the groundwater system.

Fourth, the way the alternative models are developed does not always reduce the risk of conceptual surprises. Using the Alternative Interpretation strategy, many groups will come up with what they believe to be the most likely model, e.g. Seifert et al. (2012). Using the Varying Complexity strategy, only the complexity and not conceptual ideas will be tested. It is therefore unlikely that a conceptual surprise will be found before one is surprised in both Alternative Interpretation and Varying Complexity strategy.

Last, when assigning priors to a range of models that we cannot ensure are collectively exhaustive, how do we account for unknown unknowns? The sum of prior probabilities for the ensemble of models always add up to one in the reviewed studies, thereby assuming a collectively exhaustive range of models have been defined. As discussed already, this is extremely difficult to ensure, so an approach to assign priors that accounts for unknown unknowns remains a challenge.

The Hypothesis Testing strategy seems to be the only model development strategy that can ensure the models developed are mutually exclusive. However, hypotheses might still overlap. For example, Bresciani et al. (2018) test three hypotheses to explain mountain range recharge to a basin aquifer governed either by i) mountain-front recharge, ii) mountain-block recharge or iii) both mountain-front recharge and mountain-block recharge. Some might argue that the third hypothesis overlaps to some extent with the other two, violating the mutually exclusive principle. However, only including the two first hypotheses claiming they are mutually exclusive and collectively exhaustive, would set up a false dilemma as parts of both hypotheses can be correct at the same time. It is thereby not always possible to state mutually exclusive hypotheses in hydrogeology, where the answer will be Boolean (true or false), for instance connectivity or no connectivity between aquifers (Troldborg et al., 2010). Sometimes the mutually exclusive hypothesis will have to be stated as endmembers (e.g. mountain-front recharge and mountain-block recharge) and the answer will be somewhere in between.

Guillaume et al. (2016) discuss two methods to accommodate the conceptual surprises in the model development process: Adopting adaptive management and applying models that explore the unknown. In the first approach, management plans are kept open towards change and the iterative modelling process, illustrated in Figure 2.1, is a part of the modelling plan. The second method anticipates surprises by placing fewer restrictions on what is considered possible. Stating bold hypotheses about a system, ensures that system understanding can progress (Caers, 2018). A bold hypothesis around recharge inflows from faults and deep fissures connected to an adjacent aquifer is tested by Rojas et al. (2010a). The available data did not give reason to reject either of the models to achieve an increase in system understanding, but the alternative was bold. We argue that by being forced to be bold when developing hypotheses, the risk of rejecting plausible models by omission and adopting invalid range of models is greatly reduced. However, defining bold hypotheses does not preclude rejecting plausible models by omission. Hunt and Welter (2010) suggest to use terminology that recognize the existence of these unknown unknowns by presenting results with a specification of which aspects of the model that has been considered, thereby enhancing transparency. An approach that aims at directly identifying unknown unknowns through bold hypothesis, considering the largest possible range of the conceptual uncertainty, have not been applied yet and remain a subject for further research.

2.5 HOW ARE DIFFERENT CONCEPTUALIZATIONS TESTED?

After developing a set of conceptual models, the models should be tested to establish to what degree they are consistent with the available data and knowledge (Neuman and Wierenga 2003; Refsgaard et al. 2006). Groundwater models used for safety assessment of nuclear waste repositories, for instance, have been subject of considerable validation efforts (Hassan, 2003; Rogiers et al., 2014; Tsang, 1987, 1991). Model testing and validation covers the same model evaluation process in which models are confronted with new data. However, the term validation is avoided in this review as models can never be proven correct (Konikow and Bredehoeft, 1992). Also, there is no internationally agreed definition of validation, which has led several organizations to develop their own operational definitions of validation (Perko et al., 2009). Finally, validation encourages testing to have a positive result (Oreskes et al., 1994), that is, models are not expected to be wrong. As falsification is important in order to advance our understanding of a system (Beven, 2018), the term *model testing* is preferred here.

Models are rejected if they are found to be inconsistent with data. In a Bayesian context, however, a conceptual model can never be completely rejected; its probability can only be greatly reduced. As there is a risk of eliminating models that could turn out to be good representations when new data is introduced, Guillaume et al. (2016) suggest to keep rejection decisions temporary to be able to return to otherwise excluded models. The models that are consistent with observational data are, however, only *conditionally validated* because they have not been proven to be inconsistent with data yet (Beven and Young, 2013; Oreskes et al., 1994).

Testing of conceptual models is not always done as part of the multi-model approach to groundwater modelling (Pfister and Kirchner, 2017). In Table A.2, only 30 out of 59 studies applied some form of model testing. However, model testing presents three major advantages.

First, systematically developing and testing conceptual models will allow one to explain why no other conceptual models are plausible (Neuman and Wierenga 2003), and thereby reducing the risk of adopting an invalid range of models. Through systematic documentation and rejection of conceptual models, the modelling workflow becomes transparent and traceable, potentially avoiding court cases challenging the validity of conceptual models. In the impact assessment of the Carmichael Coalmine in Queensland (Australia), available geological and hydrological data allowed for at least one other conceptualization of ecological and culturally significant springs that could potentially be impacted by the coalmine (Currell et al., 2017). However, a conceptual model leading to an acceptably low modelled impact on the springs was adopted, which lead to the approval of the mine. A systematic model development and testing approach for conceptual modelling through the multi-model approach would be able to shed light on this type of confirmation bias.

Second, model testing can lead to uncovering of unknown unknowns (Bredehoeft, 2005). Not many papers exist that actually reject all of the initial conceptual models or hypothesis about a groundwater system and come up with new plausible explanations, which renders this advantage of the model testing procedure somewhat invisible (Beven, 2018). There are, however, a few examples where models are conditionally validated after ad-hoc modifications to the model (e.g. Krabbenhoft and Anderson, 1986; Nishikawa, 1997; Woolfenden, 2008). Ad-hoc modifications are slight changes applied to a current model in order to explain conflicting data, but without falsifying the model as a whole. For example, Sanford & Buapeng (1996) developed a steady-state groundwater flow model for the Bangkok area, which was falsified by apparent groundwater ages. An ad-hoc modification that assumed groundwater velocities were higher during the last glacial maximum yielded a simulated apparent age closer to the observations, thereby conditionally validating the model with the ad-hoc modification. Ad-hoc hypotheses are sometimes criticized as they make models unfalsifiable and knowledge does not progress through modifications (Caers, 2018). However, their existence illustrates the difficulty of developing a collectively exhaustive range of models initially and model testing is imperative if we want to uncover this.

Third, Bayesian multi-model approaches benefit from allowing their prior probabilities to be updated because it dilutes the effect of the choice of priors (Rojas et al., 2009). It is here worth mentioning that most of the studies in Table A.2 that apply a Bayesian approach, update the prior probability using criteria-based weights (Section 2.6.1) while only eight studies apply a model testing procedure.

In the subsequent sections, data relevant to conceptual model testing (Section 2.5.1), steps undertaken when testing conceptual models (Section 2.5.2), and the remaining challenges within model testing (Section 2.5.3) are discussed. Table A.2 presents an overview of the model testing applied in the studies identified using the multi-model approach (Section 2.3).

2.5.1 Conceptual model testing data

Three basic requirements for the nature of the data used for model testing are typically discussed: i) it should be different from the data used for developing the conceptual models (Tarantola, 2006), ii) it should be different from the data used for calibrating the model (Neuman and Wierenga, 2003; Refsgaard et al., 2006), and iii) it should depend on the modelling purpose (Beven, 2018).

2.5.1.1 Model testing data and model building data

Tarantola (2006) distinguishes between a priori information used to develop hypotheses and observations used to test models. Post-hoc theorizing (failing to separate model development and testing data and accepting the resulting model) might lead to models being conditionally validated due to circular reasoning, e.g. the model should look this way to explain the data and the model is true because it explains the data. Another reason for keeping those two groups of data separate is to avoid underestimating conceptual uncertainty. By using geophysical SkyTEM data to both build a training image conceptual model and as soft constraint as part of a multiple-point geostatistics algorithm, He et al. (2014) demonstrated that this over-conditioning lead to an underestimation of uncertainty.

2.5.1.2 Model testing data and model calibration data

Testing data should also be different from calibration data to avoid that the conditional confirmation becomes an extension of the calibration (Neuman and Wierenga, 2003). In a review of handling geological uncertainty, Refsgaard et al. (2012) highlighted that it is possible to compensate for conceptual errors in groundwater flow models by calibrating parameters to fit the solution. The best test for any conceptualization involves comparison of model predictions to observations outside the calibration base. Cross-validation techniques, standard practice in statistical inference, are underutilised in groundwater modelling. Methodologies that minimize error variance provide some safeguard against calibration-induced acceptance of improper conceptualizations (Kohavi, 1995; Moore and Doherty, 2005; Tonkin et al., 2007).

2.5.1.3 Model testing data and the modelling objective

Refsgaard et al. (2012) further concluded that models that perform well according to one dataset might not perform well according to another dataset. This suggests that updating of prior probability should preferably be based upon the data type that the models are to make predictions about. Davis et al. (1991) argues that testing model performance outside areas relevant to the model objective can lead to rejection of models that might be fit-for-purpose. However, in many instances the data type that the models are used to make predictions, such as groundwater fluxes or water balances, may not be directly available (Jakeman et al., 2006). On the other hand, Rojas et al. (2010b) showed that by introducing more and more data in a multi-model approach, they were able to further and further discriminate between retained conceptual models, suggesting the more diverse and numerous data used for testing the more confidence in the conceptualization.

2.5.2 Conceptual model testing steps

In the previous discussion the type and nature of auxiliary data to test conceptual models were introduced. But how should such data be incorporated to undertake a conceptual model testing exercise? Neuman and Wierenga (2003) introduced a three-step workflow for testing and updating prior probability of alternative conceptual models (Table 2.1). In addition to these three steps, a fourth step, the post-audit (Anderson and Woessner, 1992) will be reviewed here.

Table 2.1. Comparison of model testing steps (pros and cons) and examples of applications in literature. The terminology of Step 1-3 is from model testing steps by Neuman and Wierenga (2003); definition of post-audit is from Anderson and Woessner (1992).

Conceptual model testing step	Pros (P) and cons (C)	Example
Step 1: “Avoid conflict with data”	Narrows down range of plausible models before conversion to mathematical model (P)	Hermans et al. (2015) tests training images for MPS against geophysical data.
Step 2: “Preliminary mathematical model testing”	Holistic test of the system (P) Parameters can compensate for conceptual error (C) Narrows down range of plausible models before complex mathematical model (P)	La Vigna et al. (2014) tests the cause of hydraulic connection between two sand aquifers against hydraulic head in a simple numerical model and can reject two out of three scenarios.
Step 3: “Confirm model”	Holistic test of the system (P) Parameters can compensate for conceptual error (C)	Parameters: Poeter and Anderson (2005) were able to reject 13 out of 61 models where the parameter distribution was wrong. State variables: Rojas et al. (2008) tested alternative conceptual models against hydraulic head and rejected two models but were unable to discriminate strongly between the rest of the models. Convergence: Poeter and Anderson (2005) rejects two models based on non-convergence.
Step 4: Post audit	Waiting time (C) Holistic test of the system (P) Parameters can compensate for conceptual error (C)	Nordqvist and Voss (1996) concluded that a supply well was in risk of contamination through a multi-model approach. After the completion of the study, increased levels of contamination were observed in the well which conditionally validated the models.

2.5.2.1 Model testing step 1

The first step in the Neuman and Wierenga (2003) guideline is referred to as “avoid conflict with data”, where the model evaluation happens before the conceptual models are converted into mathematical models. In doing so, the conceptual models can be compared quantitatively or qualitatively with data, without parameters compensating for a wrong conceptualization. Table A.2 suggests this model testing step is rarely applied, which is not necessarily true. As the evaluation of conceptual models happens outside of a numerical groundwater model, it is probably preceding the workflow in many of the studies as part of the hydrogeological investigation but not explicitly reported on. In the review by Linde et al. (2015), a workflow of corroboration and rejection is presented that focuses on the integration of geophysical data in hydrogeological modelling. For example, synthetic geophysical data may be generated from different conceptual models, and subsequently compared with observed geophysical data (Hermans et al., 2015). The prior probability of each conceptual model is then updated based on the difference between observed and simulated geophysical data. In this model testing step, however, the model evaluation does not have to be qualitative. For example, hydraulic head and electrical conductivity data may be used to distinguish between hypotheses about whether mountain front and mountain block recharge was dominating as a recharge mechanism to basin aquifers (Bresciani et al., 2018).

2.5.2.2 Model testing step 2

The second step in which data is introduced to test alternative conceptual models is called “preliminary mathematical model testing” (Meyer et al., 2007; Neuman and Wierenga 2003; Nishikawa, 1997). A similar modelling step is suggested by La Vigna et al. (2014), where for each alternative conceptual model a simple numerical model is set up and compared with testing data (hydraulic head). The advantage of applying this model testing step is that spending time on setting up complex mathematical model for poor conceptual models is avoided.

2.5.2.3 Model testing step 3

The third model testing step in Neuman and Wierenga (2003) is called “confirm model”. Here the mathematical model is set up in its most complex form. As a numerical model comprises a description of the groundwater system as a whole, all assumptions and the interaction of assumptions are tested at once. Models are then rejected either due to 1) unrealistic parameter values, 2) wrongly predicted state variables or 3) non-convergence.

Sun and Yeh (1985) showed that the optimized parameters cannot be separated from the parameter structure on which they are based on. This means if the conceptual model is incorrect, so are the estimated parameter values. Therefore, calibrated hydraulic conductivity values are often compared with “independently” measured values from pumping tests (e.g. Engelhardt et al., 2014; Harrar et al., 2003; Mechal et al., 2016; Poeter and Anderson, 2005) to check whether parameter estimates are realistic. Unfortunately, scale effects may impede direct comparison. Depending on the quality and representativeness of the data, they may or may not be able to discriminate between alternative models as was demonstrated by Engelhardt et al. (2014) and Mechal et al. (2016) for calibrated hydraulic conductivity and transmissivity values, respectively. On the other hand, in the synthetic study by Poeter and Anderson (2005), 13 out of 61 models were rejected because the calibrated hydraulic conductivity of a low-conductivity zone exceeded the conductivity of what was considered a high-conductivity zone.

Apart from comparing calibrated parameter values with observations, the predicted system variables can be compared with observations, such as hydraulic head, stream discharge, (tracer) concentrations, etc. In some multi-model studies, the number of models are limited and the comparison of simulated and observed values can happen manually. For instance, Castro and Goblet (2003) could reject all but one conceptual model by manual comparison of the direct simulation of ^{4}He concentrations with observed data. However, in cases where the lower order uncertainty is characterized within each conceptualization, automatic procedures are necessary to efficiently search for models that match field data (Rogiers et al., 2014; Rojas et al., 2010b, 2010c, 2010a; Schöniger et al., 2015a; Zeng et al., 2015). For instance, (Rojas et al., 2008) used the importance sampling technique Generalized Likelihood Uncertainty Estimation (GLUE) (Beven and Binley, 1992) to sample combinations of parameter sets and conceptual models and reject models according to an acceptance threshold for the misfit between simulated and observed model predictions.

Finally, non-convergence of the groundwater model can indicate an error in the conceptual model (Anderson et al., 2015b). The interaction of assumptions that lead to groundwater models not converging has in many studies been regarded as sufficient evidence of conceptual model invalidity (Aphale and Tonjes, 2017; Poeter and Anderson, 2005). In Rojas et al. (2008) the models that did not meet the convergence acceptance criteria were assigned a likelihood of zero, eliminating their contribution to the model ensemble predictive distribution. However, conceptual models that do not converge may potentially be valid if no effort towards making them converge is made. The effort towards making a model converge in the consensus approach will probably be larger than in the multi-model approach as there will still be other models left.

2.5.2.4 Model testing step 4

The last model testing step considered in this review is the post-audit. The post-audit is performed years after the end of the modelling process, evaluating forecasts of the model on newly collected data. Anderson and Woessner (1992) summarize some modelling studies that have used post-audits while Bredehoeft (2005) focussed on identifying the conceptual surprises that occurred in these modelling studies as a result of a post-audit. The advantage of the post-audit is that the model testing data is by default independent from the model development data, satisfying one of the basic requirements of model testing data (Section 2.5.1). However, it is inconvenient to rely on this type of model testing as there may potentially be a long waiting period from the end of the model process until new data is collected.

2.5.3 Remaining challenges

This review has shown that models can be tested in at least four different steps in the modelling process: i) as a conceptual model, ii) as a simple numerical model, iii) as a complex numerical model and iv) as a complex numerical model years after development. In each step the prior probability can be updated and sometimes models can be rejected based on lack of support by observation of state variables, parameters or because the model did not converge. Identifying suitable data for model testing remains challenging.

First, in theory the notion that testing data should be independent is sound, but in practice the separation of data is difficult. Many studies rely on ranking criteria to update the prior probability (which we will discuss in Section 2.6.1), rather than updating prior probability based on data that is independent of the model development. In using all data when developing models, it is no surprise that the models fit data. Post-hoc theorizing can easily result in undersampling of the model space (Kerr, 1998), as an initial range of plausible models will be accepted (because of circular reasoning) without looking for other plausible models. However, in many studies independent data might not be available and saving some data for the model testing process is a trade-off between being able to define a more complete model and being able to test assumptions. Cross-validation can partly address this issue during inference or calibration but will remain impractical in the conceptualization phase (model testing step 1) as biases towards existing but unavailable data might be made.

Second, in theory the data used for model testing should depend on the model objective, in order to not extrapolate when making predictions. A challenge arises when having to ensure that the model found fit-for-purpose for one dataset (e.g. hydraulic head), will also be fit-for-purpose to predict another dataset (e.g. concentrations). For example, the alternative models developed by Castro and Goblet (2003) all performed well when calibrated with hydraulic head; however, all but one model was rejected when tracer data was introduced. Sensitivity and uncertainty analysis can potentially be used to identify which parameters are relevant to the predictions and to what extent they can be constrained by the available data.

Third, the information content in the model testing data is in many studies relatively limited (e.g. Rojas et al., 2010c). The information content of model testing data relates to the amount and type of data available, but also the uncertainty of the data. For example, as discussed in relation to comparing calibrated hydraulic conductivity values to observed hydraulic conductivity values in Section 2.5.2, such comparison can be unreliable. The consequence of only limited information content in the model testing data is that discrimination among alternative models often cannot be made (Seifert et al. 2008). In addition, in a Bayesian context the consequence of limited information content in the testing data is that the prior probability will have a large influence on the posterior probability (e.g. Rojas et al., 2009).

Another challenge relates to when a model can be considered falsified. Models are groups of hypotheses rather than a single hypothesis and many other assumptions are made in groundwater models such as model code and the characterization of lower order uncertainty. The model prediction thereby depends on many interactions of independent hypotheses and assumptions. Inconsistencies between model and data should therefore not necessarily be attributed to a single hypothesis and result in the falsification of that hypothesis (Pfister and Kirchner, 2017).

To accommodate these challenges, a more systematic approach to model development and testing is needed, where parts of the available data are used only for model testing. Ideally the data selected for model testing should depend on the model objective and the information content should be large enough to discriminate between models. There is thereby an opportunity for systematic (quantitative or qualitative) assessment prior to study (i) which aspects of the model will be relevant to the objectives and (ii) what data are needed to distinguish between hypotheses.

2.6 HOW ARE DIFFERENT CONCEPTUALIZATIONS USED FOR PREDICTIONS?

What has emerged from several of the studies so far in this review is that multiple plausible models may coexist for a given study area. So, how are predictions made with multiple models? For some studies (e.g. Foglia et al., 2013), one model (the most likely based on the highest support in data) is selected for predictive purposes (Section 2.6.1), while other studies (e.g. Tsai and Li, 2008) focus on ensemble predictions based on all plausible models (Section 2.6.2). A modelling step that receives increasing attention in the literature is the identification of additional data needs in order to be able to discriminate between the alternative conceptual models (e.g. Kikuchi et al., 2015) (Section 2.6.3). The last four columns in Table A.2 present an overview of approaches being adopted when making predictions with multiple models. As mentioned in the introduction, several literature reviews (Diks and Vrugt, 2010; Schöniger et al., 2014; Singh et al., 2010) have already focussed on the model prediction and evaluation aspect of the multi-model approach. It is therefore not the aim to give a comprehensive review here, but to give a general overview of the most often applied approaches and instead focus on how the model development approach (discussed in Section 2.4) affects the predictions.

2.6.1 Model weighing and selection techniques

Model weighing and selection techniques rank models according to how well they fit data, where the models with the lowest rank or weight have least support in the data. The purpose of ranking is to select the “best” model, but for many of the studies in Table A.2 ranking also provides weights for a model averaging technique (Section 2.6.2). For an excellent review of model selection techniques the reader is referred to Schöniger et al. (2014).

In selecting between models, two principles often receive attention: The Principle of Parsimony (favouring the simplest model) and The Principle of Maximum Likelihood (favouring the model that gives the highest chance to facts we have observed). However, the Principle of Consistency (favouring models that do not contradict any effects we know) is even more important to consider when choosing between models (Martinez and Gupta, 2011). The most commonly applied ranking techniques in the analysed studies in Table A.2. are the Information Criteria, including Akaike’s Information Criterion (AIC) (Akaike, 1973), corrected AIC (AICc) (Sugiura 1978; Hurvich and Tsai 1989), Bayesian Information Criterion (BIC) (Schwarz, 1978) and Kashyap Information Criterion (KIC) (Neuman, 2003) and GLUE. The ranking from the information criteria depends on an error term representing model fit to observations and a penalty term that penalizes model complexity. In GLUE the ranking is only based on an error term.

2.6.2 Model averaging techniques

Model averaging techniques seek to summarize the results from the multiple model approach into an optimal prediction and a single measure of the total uncertainty by averaging the posterior distributions (Raftery et al., 2005). This posterior is obtained through an averaging approach that weighs the different model predictions according to the weight they obtained from the testing or ranking, combined with a prior probability of the individual models. For excellent summaries of model averaging techniques the reader is referred to Diks and Vrugt (2010) and Singh et al. (2010).

The most commonly applied approach to averaging predictions of conceptually different hydrogeological models is Bayesian Model Averaging (BMA) (Hoeting et al., 1999). The averaged predictions from multiple models have been shown to be more robust and less biased than the prediction from a single model (Vrugt and Robinson, 2007). Furthermore, they produce a more realistic and reliable description of the predictive uncertainty (Rojas et al., 2010a).

The Bayesian model evidence is sometimes approximated with the information criteria to reduce computational effort constituting the Maximum Likelihood BMA (MLBMA) approach suggested by Neuman (2003). Given many of the information criteria are developed as model selection criteria, they tend to assign a large weight to only a few models (e.g. Nettasana, 2012; Rojas et al., 2010c; Ye et al., 2010), which is the main drawback of the MLBMA approach. This leads to the introduction of a statistical scaling factor to the information criteria (Tsai and Li 2008), leading to a flatter weight distribution among the alternative models.

One of the disadvantages of the averaging procedures is that the system details of how each conceptual model affects the prediction, is lost (Gupta et al., 2012). To solve this problem, Tsai and Elshall (2013) suggested the hierarchal BMA (H-BMA) approach where the individual conceptual model components are evaluated through a BMA tree. In the BMA tree model components are organized at separate levels and the contribution of uncertainty of each aspect to the total uncertainty is quantified. By separating the uncertain model components in a BMA tree, the different aspects can be prioritized and provide an understanding of the uncertainty propagation through each uncertain aspect in the conceptual model.

2.6.3 Identify additional data needs

Refining the prediction made by multiple models may sometimes be necessary in order to decrease the range of model predictions. Considering too many conceptual models, one may lose the purpose of model development because it indicates high model prediction uncertainty (Bredehoeft, 2005; Højberg and Refsgaard, 2005). Therefore, some studies have focussed on identifying additional data needs that could potentially discriminate between alternative conceptual models to reduce conceptual uncertainty (e.g. Kikuchi et al., 2015; Pham and Tsai, 2015, 2016). The goal of collecting new data is not to confirm existing conceptual models, but to be able to discriminate between them.

Kikuchi et al. (2015) offers a short review of optimal design studies in hydrogeology that attempt to identify the optimal measurement sets for monitoring networks to maximize a data utility function. For a few studies conceptual model discrimination is the design objective (Knopman et al., 1991; Knopman and Voss, 1988, 1989; Usunoff et al., 1992; Yakirevich et al., 2013), but this approach has yet not received much attention in hydrogeology according to Kikuchi et al. (2015).

Identifying additional data needs will guide the post audit activity (Section 2.5.2.4) and the use of these data for model testing will ensure the data is independent from the model development data.

2.6.4 Remaining challenges

This review shows that current studies often either used criteria-based weights, either to identify the most plausible models or to provide weights for a model averaging technique. The current methods are generally limited by what is attainable through the model development approach. The main limitations and thereby consequences of the model development approach for current methods on making predictions with multiple conceptualizations are discussed next.

First, we can never make sure that we have developed a collectively exhaustive range of conceptual models (e.g. Ferré, 2017; Hunt and Welter, 2010; Nearing and Gupta, 2018) (as discussed in Section 2.4) but the prediction methods and the approaches in identifying additional data types rely on this. Undersampling the model space will lead to underestimation of the prediction uncertainty in the model averaging approaches. Furthermore, by focussing the collection of additional data on data that can discriminate between currently known conceptualizations, it is assumed that we already know all plausible conceptualizations. A challenge remains in directing additional data collection towards uncovering unknown unknown plausible conceptual models.

Second, we can never make sure that the adopted range of models developed is valid (Type II error) (e.g Nearing and Gupta, 2018) but both the BMA and the criteria-based model weighing techniques rely on the best approximation of reality being in the ensemble. In the model selection approaches we can therefore never make sure that the best approximation of reality is selected as it will always be conditional on the developed range of models. In the model averaging approaches, adopting an invalid range of models leads to biased predictions, which remains a challenge.

Third, in BMA it is assumed that models are mutually exclusive, so that some predictions are not given a higher weight following almost identical models give similar predictions. Not having mutually exclusive models gives a false sense of confidence in the modelling results, as a large number of alternative models considered will give the impression that a large range of the model space has been uncovered.

Fourth, the criteria-based model weighing techniques rely only on the Principle of Parsimony and the Principle of Maximum Likelihood, while the Principle of Consistency is disregarded through calibration. Through the calibration step the model is trained to compensate for a possible conceptual error through biased parameters (Refsgaard et al., 2012; White et al., 2014) and the Principle of Consistency is therefore not taken into account. Criteria-based model weighing techniques use the same data twice in the modelling process, which as discussed in Section 2.5.1, leads to circular reasoning giving a false confidence in the result. Also, inconsistent assumptions in the conceptual model cannot be identified without introducing new data, but in the criteria-based model weighing techniques, models are readily rejected through zero-weight as they tend to inflate the weights of a few best models (e.g. Ye et al., 2010). The models that best compensate for conceptual errors through biased parameters are then combined to make predictions through model averaging, where it is claimed that conceptual model uncertainty is considered. However, given the biased parameters of the models, circular reasoning and rejection of plausible models, this result may be both biased and over-conservative.

Last, the model averaging techniques assume that a single result is valid, however if the range of plausible model are mutually exclusive, they might lead to distinctly different predictions. One model might have a distinctly different prediction than the ensemble average or the probability mass may concentrate in multiple areas. This is the case for the synthetic example in the study by Kikuchi et al. (2015), where the spring flow depletion prediction is bimodal. In this case the average prediction is an outlier to where the probability mass is concentrated. The average prediction of an ensemble, especially bi- or multi-modally distributed ensembles, may not be a valid model outcome (Winter and Nychka, 2010). It is therefore preferable to summarise ensembles through more robust metrics, such as percentiles (e.g. 5th, 50th and 95th) as these will always be actual results made by a model.

Suggestions on solving the remaining challenges in relation to populating the model space (first, second, third point) has already been discussed in Section 2.4.5. The challenges mentioned in the remaining two points occur because of the reliance on methods that assume a single best model can be found. A way forward to accommodate these challenges could be full probabilistic approaches. Trans-dimensional inference methods have been applied in geophysics (e.g. Ray and Key, 2012) and reservoir geology (e.g. Sambridge et al., 2006) for similar problems. In these approaches, e.g. reversible jump Markov Chain Monte Carlo (Green, 1995), sampling occurs within the same dimension (conceptual model), but also between dimensions (conceptual models) exploring both the conceptual model space and the parameter space.

2.7 CONCLUSION

A review of 59 studies applying the multi-model approach for hydrogeological conceptual model development, has shown the following:

1. A significant linguistic uncertainty still exists of what is considered conceptual uncertainty. There is a need for more consistent terminology.
2. Current studies in conceptual model uncertainty often only focus on a single or limited set of conceptualization issues. There is a need for a systematic conceptualization approach to ensure all aspects of conceptualization are covered and documented.
3. Current studies rarely consider the objective of the model before developing alternative models for the multi-model approach. The objective of the model should have an influence on both the model development and the data used for model testing.
4. For each conceptual issue identified, alternative conceptual models should be formulated as hypotheses which, at least in theory, can be refuted. Hypothesis testing, especially bold hypothesis testing, is essential to increase system understanding and avoiding conceptual surprises.

5. In Bayesian inference with multiple models, informed priors are recommended, especially if the information content in the hypothesis testing data is low.
6. The current multi-model prediction methods assume that there is a single outcome of the modelling process and that the developed models are mutually exclusive and collectively exhaustive. Presenting results requires a shift in mentality towards presenting ranges and acknowledging that unknown unknowns exist.

The multi-model approach is superior to the consensus approach as it is transparent and accounts for conceptual uncertainty. However, to benefit fully from the multi-model approach, challenges remain in being more systematic in regard to both developing and testing alternative models.

Chapter 3: Bayesian Hypothesis Testing in a Stochastic Water Balance Model

Trine Enemark, Luk JM Peeters, Dirk Mallants, Okke Batelaan, Andrew P. Valentine, Malcolm Sambridge

Published in *Water*[2]

[2] Enemark, Trine, Peeters, L. J. M., Mallants, D., Batelaan, O., Valentine, A. P., & Sambridge, M. (2019). Hydrogeological Bayesian Hypothesis Testing through Trans-Dimensional Sampling of a Stochastic Water Balance Model. Water, 11(1463). https://doi.org/10.3390/w11071463

3.1 ABSTRACT

Conceptual uncertainty is considered one of the major sources of uncertainty in groundwater flow modelling. In this regard, hypothesis testing is essential to increase system understanding by refuting alternative conceptual models. Often a stepwise approach, with respect to complexity, is promoted but hypothesis testing of simple groundwater models is rarely applied. We present an approach to model-based Bayesian hypothesis testing in a simple groundwater balance model, which involves optimization of a model in function of both parameter values and conceptual model through trans-dimensional sampling. We apply the methodology to the Wildman River Area, Northern Territory, Australia, where we set up 32 different conceptual models. A factorial approach to conceptual model development allows for direct attribution of differences in performance to individual uncertain components of the conceptual model. The method provides a screening tool for prioritizing research efforts while also giving more confidence to the predicted water balance compared to a deterministic water balance solution. We show that the testing of alternative conceptual models can be done efficiently with a simple additive and linear groundwater balance model and is best done relatively early in the groundwater modelling workflow.

3.2 INTRODUCTION

The conceptualization of a groundwater flow problem is considered one of the major sources of uncertainty in groundwater flow modelling (Enemark et al., 2019a; Gupta et al., 2012). Conceptual uncertainty stems from the fact that the available data more often than not will fit more than one conceptual understanding (Bredehoeft, 2005). When dealing with conceptual uncertainty, hypothesis testing is essential to increase system understanding by refuting alternative conceptual models (Beven, 2018).

The question we ask of the hypothesis testing exercise is framed by the model development approach. In practice, individual conceptual hypotheses cannot be tested through model based hypothesis testing; only collections of hypotheses can be tested (Nearing and Gupta, 2018; Oreskes et al., 1994). That is, if a hypothesis cannot be falsified in a model, it is only conditionally validated given the assumptions in other parts of the model. The challenge is to develop alternative models so that differences in performance can be attributed to individual hypotheses. For exploratory purposes, model development should aim at maximizing the difference between alternative models in order to gain most information from a potential model rejection (Caers, 2018; Guillaume et al., 2016).

One branch of hypothesis testing is based on Bayesian probability theory. In Bayesian hypothesis testing, a prior belief about the suitability of a conceptual model is updated to a posterior belief by evaluating the model performance against data. The performance of alternative models are then compared in order to quantitatively rank and potentially reject hypotheses based on the so-called Bayes factor (Jeffreys, 1939; Kass and Raftery, 1995). Fields of application in hydrogeology include groundwater modelling (Rojas et al., 2010c, 2010a), hydrogeophysics (Brunetti et al., 2017; Hermans et al., 2015) and solute transport modelling (Thomsen et al., 2016; Troldborg et al., 2010). In many applications, the data is not sufficient to allow for discrimination between the models, in which case Bayesian model averaging is often applied where model predictions are weighed according to their performance against data (Höge et al., 2019).

In hydrogeology, conceptual models are often tested in mathematical models, (e.g. Dausman et al., 2010; Remson et al., 1980). Since a model comprises the description of a system as a whole, all assumptions and the interaction of assumptions are tested at once. Also, data that does not directly relate to the conceptually uncertain feature can be integrated, because of the holistic testing of the system.

A stepwise approach in regards to complexity to groundwater flow modelling and hypothesis testing is often promoted (Haitjema, 1995; Neuman and Wierenga, 2003). The simplicity of a model is in this paper defined in terms of setup- and run-time. In a stepwise approach, complexity is gradually built up, and involves testing the models in each step to better understand the relative importance of various assumptions. This is opposed to starting with a complex model where all known processes and structural aspects are incorporated "because they exist, not because they matter" (Haitjema, 1995).

Although there seems to be consensus that testing simple models is advantageous, most model based hypothesis testing in hydrogeology happens in complex models. We argue that there is a need to also test models as early as possible with models being as simple as possible for at least four reasons. First, simple models can offer insight into system understanding that can be obscured in more complex models (Haitjema, 2006; Hunt and Zheng, 2012). Second, testing models early in the workflow enables identification of important sources of uncertainty and knowledge gaps to help prioritize research efforts, including data collection (Turnadge et al., 2018b). Third, pragmatic constraints on time and budget limit the number of models that can be tested and fewer models are tested when they are more complex (Refsgaard et al., 2012). Finally, rigorous model testing allows to identify conceptual surprises early in the research effort, rather than detecting them after modelling has been completed. A stepwise approach to groundwater modelling is especially important when the multi-model approach is adopted; this is still mainly an academic exercise as developing several conceptual models rather than a single is a time-consuming task.

One of the most widely applied and simplest approaches to represent a groundwater system is through an additive groundwater system water balance (Barnett et al., 2012; Dassargues, 2018). Water balances are systematic records of the water fluxes going into and out of a groundwater system, and how these fluxes affect the stored volumes of water. The estimation of uncertainties in water balance components have been the topic of several studies (e.g. Baalousha, 2009; Sebok et al., 2016; Thompson et al., 2017), but conceptualization issues are rarely considered. By applying different conceptualizations, the number of parameters describing the water balance is variable. A trans-dimensional (Green, 2003) inverse problem is one where the dimension of the parameter space, and not just values of the parameters, is a variable to be solved for. Within the geosciences, trans-dimensional sampling has most often been applied in geophysics (e.g. Malinverno and Leaney, 2000), but has gained ground in hydrology in recent years (e.g. Jiménez et al., 2016; Mondal et al., 2010; Somogyvari et al., 2017). In this paper we apply trans-dimensional sampling to a stochastic water balance in order to test conceptually uncertain components in the water balance problem.

The aim of this study is to (1) develop a model development framework to test alternative conceptual models in a Bayesian framework, (2) apply it to a simple groundwater balance model and (3) evaluate if there is sufficient information in a water balance to gain insight on the conceptualization of a groundwater system. We apply the methodology to the Wildman River Area in the Northern Territory, Australia.

3.3 MATERIALS AND METHODS

A simplified representation of the applied methodology is illustrated in Figure 3.1. In the stochastic water balance, different components, represented by different coloured circles, are included or excluded in the water balance equation in order to represent conceptual uncertainty. In each parameter realization, the magnitude of the components are varied in order to represent aleatory uncertainty (i.e. uncertainty that can be modelled stochastically), represented by the size of the circle in Figure 3.1a. The likelihood of each of these parameter and model realizations is based on the error of the water balance and a Metropolis-Hastings sampler (Hastings, 1970; Metropolis and Ulam, 1949) is applied over these likelihoods (Figure 3.1b) in order to estimate the probability of the different plausible models. The more likely a model is, the more often the sampling algorithm will visit the model. The posterior probability of a model is the number of accepted visits that is based on an acceptance probability. An inter-comparison of the posterior model probabilities reveals whether some models are preferred over others based on predefined threshold values of the Bayes factor. More details about the workflow illustrated in Figure 3.1, will be provided in Section 3.3.4.

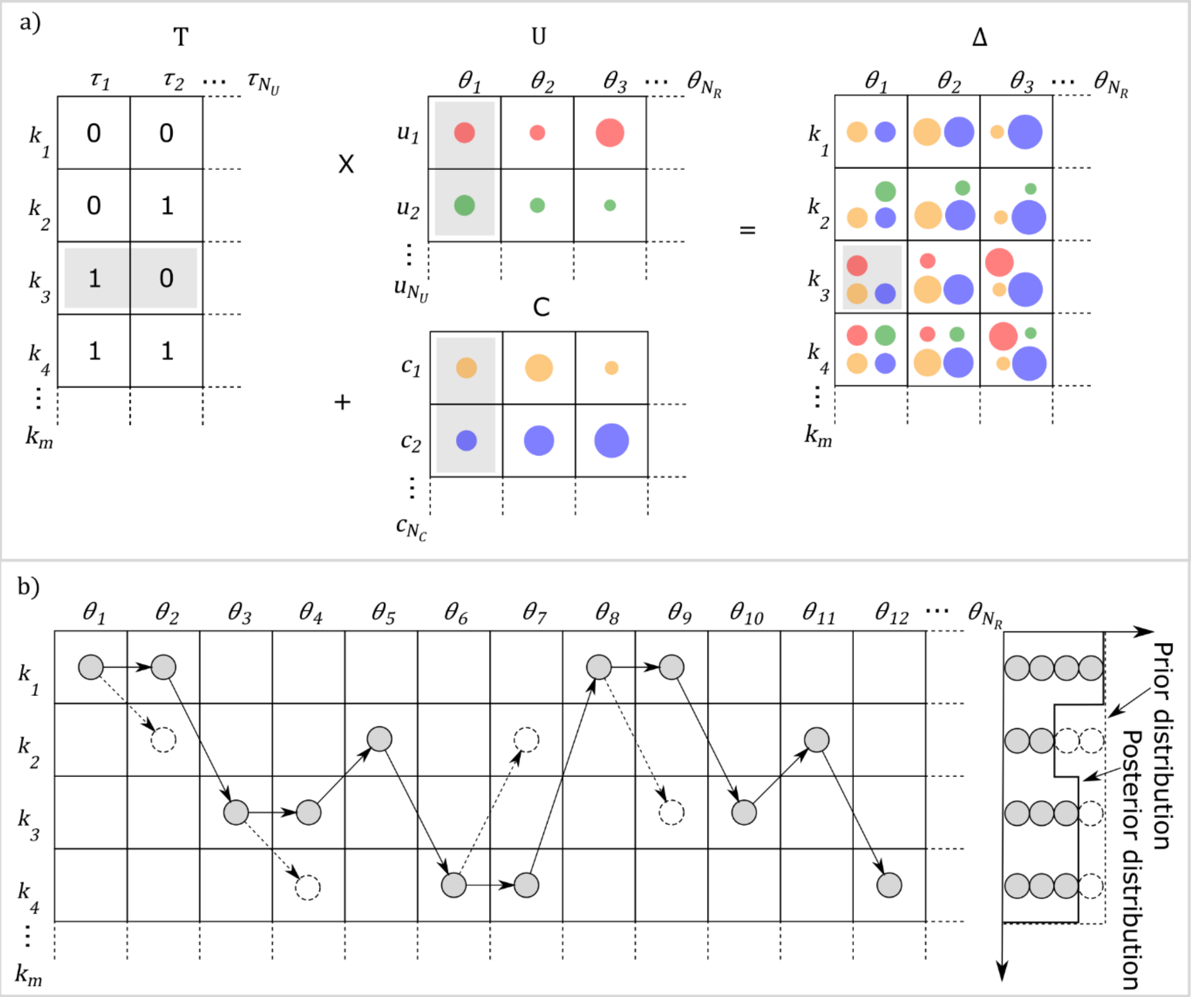

Figure 3.1. a) Method for calculation of the water balance error, Δ. The colors of the circles represent different system subcomponents, C, while the size of the circles represents the magnitude of the subcomponent in each realization, θ_i with $i \in \{1, \dots, N_R\}$ where N_R is the number of variable realizations. The combination of τ_h with $h \in \{1, \dots, N_U\}$, where N_U is the number of uncertain model components, in the rows of T, constitutes a model, k_j with $j \in \{1, \dots, m\}$ where $m = 2^{N_U}$. The uncertain model components in U includes N_U components, while the certain model components in C includes N_C components. b) Metropolis-Hastings sampling of the water balance realizations. The circles and arrows represent that are accepted or rejected based on the acceptance criteria, α. The dotted circles and arrows represent proposed moves that were not accepted. Figure 3.1b) modified from (Lee et al., 2015).

In the following section, a detailed description of the methodology is presented. First, we propose an alternative model development methodology, and introduce the Bayesian inference and interpretation methods. The methodology described in these sections are entirely generic and therefore applicable to any type of model. Last, we describe the setup of the groundwater balance that facilitates effortless evaluation of numerous model realization.

3.3.1 Model development method

Conceptual models can generally be decomposed into a collection of hypotheses that describe the conceptual physical structure and the conceptual process structure (Gupta et al., 2012).

Note that the term conceptual model comprises a collection of hypotheses, while the term hypothesis concerns individual uncertain components in the conceptual model. More often than not, uncertain components will exist in the conceptual model, either to do with the geometry (e.g. hydrostratigraphy) or processes (i.e. boundary conditions) in the groundwater system. We suggest defining hypotheses for each conceptually uncertain component in the conceptual model in the following manner:

- H_0: The process/geometry does not matter for the prediction of interest.
- H_A: The process/geometry matters for the prediction of interest.

In the null hypothesis (H_0) the uncertain component is excluded, while the alternative hypothesis (H_A) will include the uncertain component. Although we apply a Bayesian hypothesis testing framework, we borrow the null and alternative hypothesis terminology from null hypothesis significance testing to illustrate that the two hypotheses are mutually exclusive. This framework ensures models are mutually exclusive, so that the probability of the null hypothesis and the alternative hypothesis for a single uncertain component adds up to one. Making a statement about whether the conceptually uncertain component matters for the prediction of interest rather than whether it is present or not, places emphasis on the objective of the modelling exercise and makes the hypotheses easier to disprove.

In a mathematical model, we can add an extra parameter to represent the conceptual uncertainty. We can turn an uncertain conceptual components on or off in function of τ_h with $h \in \{1, \ldots, m\}$ that takes a value of either 0 or 1. Here $\tau_h = 0$ represents the null hypothesis and $\tau_h = 1$ represents the alternative hypothesis. For the number of uncertain components, $N_{U,}$ the number of possible combinations of the null and the alternative hypotheses is $m = 2^{N_U}$. All possible models can be defined in matrix $\mathbf{T}$, where each row represents an individual conceptual model, k_j with $j \in \{1, \ldots, m\}$ (Figure 3.1a):

$$\mathrm{T} = \begin{array}{cc} \begin{bmatrix} \tau_{1,1} & \tau_{1,2} & \cdots & \tau_{1,N_U} \\ \tau_{2,1} & \tau_{2,2} & \cdots & \tau_{2,N_U} \\ \vdots & \vdots & \ddots & \vdots \\ \tau_{\mathrm{m},1} & \tau_{\mathrm{m},2} & \cdots & \tau_{\mathrm{m},N_U} \end{bmatrix} & \begin{matrix} k_1 \\ k_2 \\ \vdots \\ k_{\mathrm{m}} \end{matrix} \end{array} \tag{3.1}$$

By developing alternative models with a factorial design (Fisher, 1935) of uncertain model components, the difference in model performance can be attributed to individual uncertain model components.

In hydrogeology, similar approaches have previously been presented, (e.g. Aphale and Tonjes, 2017; Pham and Tsai, 2016; Troldborg et al., 2010). This factorial approach is in contrast to approaches where alternative model development hypotheses are grouped in alternative models (Højberg and Refsgaard, 2005; Seifert et al., 2012; Ye et al., 2008a), thus limiting the information that can be gained from the modelling exercise. Hierarchical Bayesian Model Averaging (HBMA) (Chitsazan et al., 2015; Tsai and Elshall, 2013), presents a similar approach that also aims at separating and quantifying the contribution of uncertain model components to prioritize and provide an understanding of a conceptual model. The main difference between the model building approach in HBMA and the one presented here, is that in HBMA the hypotheses in are not necessarily mutually exclusive.

3.3.2 Bayesian inference framework

The goal of the Bayesian inference is to compute posterior probabilities for parameters and hypotheses based on data. When the dimension of the parameter vector is one of the unknowns, the joint posterior probability, $p(k, \theta_k|Y)$ of the model indicator k and a parameter vector, θ_k given the data, Y, becomes the basis of the inference and the problem can be categorized as trans-dimensional. The inference starts from a prior probability of models, $p(k)$, and a prior probability of parameters $p(\theta_k|k)$, for each model, k. The prior is linked to the posterior through a likelihood function, $p(Y|k, \theta_k)$:

$$p(k, \theta_k|Y) = p(Y|k, \theta_k)\, p(\theta_k|k) \tag{3.2}$$

The likelihood function describes the probability of the observed data given the model. To approximate the posterior probability we apply Markov Chain Monte Carlo simulation where each iteration consists of moving from the current parameter vector θ^p to the proposed parameter vector θ^q and from the current model k to the proposed model, k' (Figure 3.1b). If the proposal distribution is symmetric, each iteration step can be accepted with an acceptance probability, α, of (Sambridge et al., 2006):

$$\alpha = \min\left\{1, \frac{p(Y|\theta^q, k')p(k')}{p(Y|\theta^p, k)p(k)}\right\} \tag{3.3}$$

In practice, models are accepted if a uniform random number between 0 and 1 is lower than the acceptance probability, α. If the acceptance probability is lower than this random number between 0 and 1, the proposed model position (θ^q, k') is rejected and the algorithm stays at position (θ^p, k).

3.3.3 Interpretation

The Bayesian hypothesis testing problem involves interpretation of the marginal evidence for each model, $p(Y, k)$. The marginal evidence is obtained by integrating the posterior over all plausible model parameters and describes how well the model explains data taking all plausible parameter combinations into account.

For pairwise comparison of the evidence provided by data for models, the Bayes factor can be calculated. When the probabilities are converted to an odds scale (odds = probability/(1 – probability), the Bayes factor is defined as (Kass and Raftery, 1995):

$$B_{1,2} = \frac{p(k_1|Y)p(k_2)}{p(k_2|Y)p(k_1)} \tag{3.4}$$

where the two indices, 1 and 2, are used here as a simple example to indicate two mutually exclusive models. When prior probabilities are uniform, the Bayes factor is equal to the ratio of posterior probabilities. To ease interpretation, Kass and Raftery (1995) applied descriptions of the evidence provided for intervals of the Bayesian factor (Table 3.1). A Bayes factor below 1 shows support for the model in the denominator (k_2). These thresholds can be used as a guideline for decisions about model selection or rejection, however the interpretation may depend on context.

Table 3.1. Interpretation of level of support for model k_1 or k_2 from Bayes factors as defined in equation (3.4), with k_1 in the numerator and k_2 in the denominator. When the probability of k_1 and k_2 is 1, the ranges of Bayes factor correspond to the probabilities in the "Probabilities" column.

Bayes Factor	Probabilities	Description
< 0.005	< 0.075	**Decisive** support for **k$_2$**
0.005 – 0.05	0.075 – 0.182	**Strong** support for **k$_2$**
0.05 – 0.3	0.182 – 0.366	Substantial support for k$_2$
0.3 – 3	0.366 – 0.634	Inconclusive, no support for either k$_1$ or k$_2$
3 – 20	0.634 – 0.818	Substantial support for k$_1$
20 – 150	0.818 – 0.925	**Strong** support for **k$_1$**
> 150	> 0.925	**Decisive** support for **k$_1$**

3.3.4 Water balance model

The additive groundwater balance approach is based on the conservation of mass principle subtracting the water flowing out of the aquifer from the water flowing into the aquifer over a specified time period. We use positive numbers for water going into the aquifer and negative numbers for water going out of the aquifer. In this study we will independently identify the contribution of each component in the water balance, so that no component will have to be estimated from the residual. Not assuming perfect water balance closure is a requirement for the suggested method as the water balance error will be used to assess the likelihood of the model. The groundwater balance can be written:

$$Q_{\text{Input}} = -Q_{\text{Output}} \pm \Delta S \pm \delta \tag{3.5}$$

Where Q_{Input} is the quantity of water entering the watershed (e.g. recharge, lateral recharge, river recharge); Q_{Output} is the quantity of water leaving the watershed (e.g. lateral discharge, river discharge, evapotranspiration); ΔS is the change in storage over the specified period of time and δ is the error term that represents the remaining error in the water balance. The water balance may contain many subcomponents within the input and output component and a more general definition is therefore:

$$\delta = \pm c_1 \pm c_2 \ldots \pm c_{\mathrm{N}_c} \tag{3.6}$$

Where each c is a subcomponent in the water balance, either positive or negative, depending on whether water is flowing into or out of the aquifer and N_c is the number of known subcomponents of the water balance. The different subcomponents are represented by different coloured circles in Figure 3.1a.

Conceptual uncertainty in a water balance arises when some or all parts of the water balance components are hypothesized to exist, but not known to exist. We set up several hypotheses for the uncertain water balance components using the method described in Section 3.3.1:

- H_0: Water balance component does not matter for the prediction of interest.
- H_A: Water balance component matters for the prediction of interest.

By applying the τ parameter to capture conceptual uncertainty as described in Section 3.3.1, the water balance can then be expanded to:

$$\delta = \pm c_1 \pm c_2 \ldots \pm c_{\mathrm{N}_C} \pm \tau_1 u_1 \pm \mathrm{\tau}_2 u_2 \ldots \pm \tau_{\mathrm{N}_U} u_{\mathrm{N}_U} \tag{3.7}$$

Where each u represents a conceptually uncertain subcomponent who are all associated with a τ value and N_U is the number of conceptually uncertain components.

The uncertain components can be organized in a matrix **U** (Figure 3.1a) where the number of rows equals the number of uncertain components, N_U, while the certain components are organized in the matrix **C,** where the number of rows equals the number of certain components, N_C:

$$U = \begin{bmatrix} u_{1,1} & u_{1,2} & \cdots & u_{1,N_R} \\ u_{2,1} & u_{2,2} & \cdots & u_{2,N_R} \\ \vdots & \vdots & \ddots & \vdots \\ u_{N_U,1} & u_{N_U,1} & \cdots & u_{N_U,N_R} \end{bmatrix}, \qquad C = \begin{bmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,N_R} \\ c_{2,1} & c_{2,2} & \cdots & c_{2,N_R} \\ \vdots & \vdots & \ddots & \vdots \\ c_{N_C,1} & c_{N_C,1} & \cdots & c_{N_C,N_R} \end{bmatrix} \tag{3.8}$$

In both matrices each column represents an individual parameter realization, where N_R is the number of parameter realizations. In each realization the value of the subcomponent will be different, illustrated by the different sizes of the circles in Figure 3.1a. The magnitude of the subcomponents is modelled stochastically by drawing the value from a predefined prior distribution.

By computing the dot product of **T** and **C** and adding **U,** all models can be simulated at the same time for all parameter vectors. This will yield a matrix **Δ** in which each row presents realizations within each individual conceptual model, k_j with $j \in \{1, \dots, m\}$ and each column represents an individual realization based on different parameter vectors θ_i with $i \in \{1, \dots, N_R\}$ (Figure 3.1a):

$$\Delta = T \cdot C + U = \begin{array}{cc} \begin{bmatrix} \delta_{1,1} & \delta_{1,2} & \cdots & \delta_{1,N_R} \\ \delta_{2,1} & \delta_{2,2} & \cdots & \delta_{2,N_R} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{\mathrm{p},1} & \delta_{p,2} & \cdots & \delta_{\mathrm{p},N_R} \end{bmatrix} & \begin{matrix} k_1 \\ k_2 \\ \vdots \\ k_p \end{matrix} \\ \begin{matrix} \theta_{\mathrm{p},1} & \theta_{\mathrm{p},2} & \cdots & \theta_{\mathrm{p},N_R} \end{matrix} & \end{array} \tag{3.9}$$

In Figure 3.1a, a zero in the **T** matrix will exclude the subcomponent in matrix **Δ** while a one will include the subcomponent illustrated by the presence and absence of the coloured circles in matrix **Δ**. This setup of the water balance problem takes advantage of parallelization and vectorization, enabling millions of random realizations to be realized per second. The metropolis sampling algorithm described in Section 3.3.2 can be defined on top of the already generated realizations by sequentially stepping through the columns of the matrix (Figure 3.1b). The proposed model k' is randomly chosen with weights according to the likelihoods of the different models based on the same parameter vector.

Likelihoods ($p(Y|k, \theta_k)$) can be computed based on the error of the water balance in matrix $\mathbf{\Delta}$. The water balance error δ is scaled to the magnitude of the input to the water balance (Q_{Input}) to get a relative error. By using the relative error, a larger error is accepted for water balances with large water balance components, and smaller error for water balances with small water balance components. All priors will be constrained by data, so that all samples of the magnitude of the water balance components are considered plausible.

The relative error term is assumed to be normally distributed with a mean of 0 and a standard deviation σ :

$$p(Y|k, \theta_k) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{\left(\frac{\delta}{Q_{Input}} - 0\right)^2}{2\sigma^2} \qquad (3.10)$$

The observation error is here assumed to be captured by the parameters set of the prior probabilities of parameter values and σ is therefore only related to the remaining conceptual error in the model representing the unknown unknown conceptual components. The standard deviation of the water balance error directly controls the acceptance rate of the water balance realizations; as σ increases, more realizations will be accepted. It should be stated that as for all error formulas, errors of opposite sign will cancel out, reducing the overall error, and a conceptual model with many unknown unknowns might therefore perform well in this framework.

As the magnitude and number of unknown unknowns in the model is unknown, there is no way of determining the value of σ. The robustness of model ranking as defined by Schöniger et al. (2015b), has been evaluated in a sensitivity analysis by varying the standard deviation of the relative error of the water balance between 0 % and 10 %. To avoid making wrong model selection decisions based on the results, we have selected a standard deviation of the model error that is least decisive. That is, the difference between the probability of the null and the alternative hypotheses are closest to each other. The least decisive value is the most conservative choice when the model objective is to differentiate between models. From the sensitivity analysis we determined that the most conservative value is $\sigma = 2.5$ %.

3.4 CASE STUDY

The Wildman River Area (Northern Territory, Australia) was chosen as case study area (Figure 3.2). The area covers about 400 km^2 in the northern part of Northern Territory, next to the Kakadu National Park, between latitudes -12.77 and -12.47 and between longitudes 131.7 and 131.93.

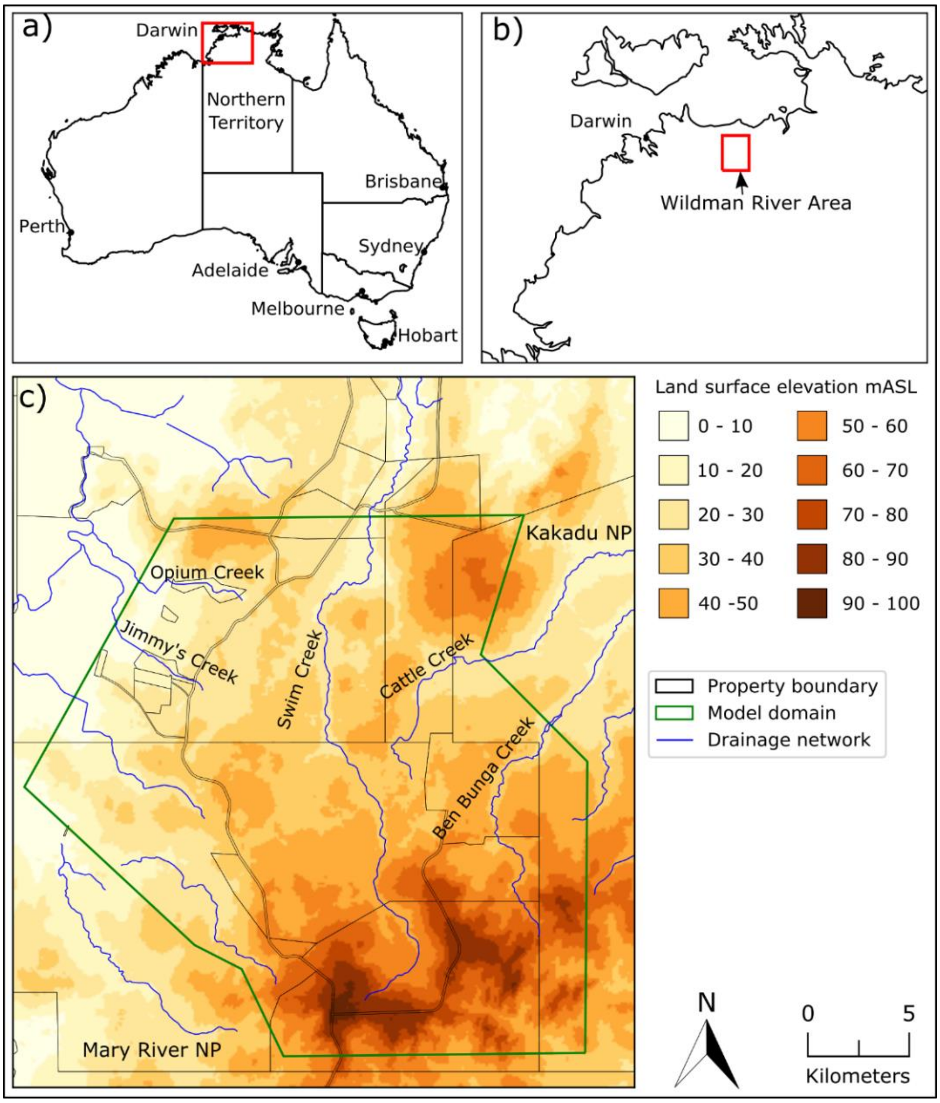

Figure 3.2. The study area is a part of the Wildman River Area located between Mary River and Kakadu National Park in Northern Territory, Australia.

The topography in the catchment is relatively flat, and ranges from 100 m ASL in the south to 20 m ASL in the north (Figure 3.2). Most of the land use is limited to cattle grazing and conservation. Mary-River National Park lies in the south-west of the area and Kakadu National Park to the east. The area has a tropical climate with about 98% of annual precipitation (1450 mm) occurring in the wet season, December to April (Turnadge et al., 2018a).

The basement geology in the region mainly consists of three units of the Mount Partridge Group which is a part of the Pine Creek Orogen: the Wildman Siltstone, the Koolpinyah Dolostone and the Mundogie Sandstone. The basement is tight to isoclinally folded with a strike of between 180 and 200 degrees. Unconsolidated sediments deposited in the Money Shoal Basin are unconformably overlaying the basement. These sediments are flat laying but repeated incision and infill through the Cenozoic is thought to have created two southwest-northeast oriented palaeovalleys. Tickell and Zaar (2017) referred to these sediments as Mesozoic-Cenozoic (Mz/Cz) sediments, as ambiguity surrounds their age.

The main aquifers consist of a semi-confined Mz/Cz sand aquifer and a confined dolostone aquifer that are assumed to be connected, supported by similarity in hydrogeochemistry (Tickell and Zaar, 2017). Very limited topographical and piezometric data in the area suggests that the main flow direction is vertical with a very slow horizontal component.

Two major investigations have been carried out in the area by the Northern Territory Department of Environment and Natural Resources in relation to a water resource assessment for the area Tickell and Zaar (2017) and CSIRO as part of the Northern Australia Water Resources Assessment (Turnadge et al., 2018a, 2018c). (Tickell and Zaar, 2017) provided a first-order assessment of the regional-scale groundwater balance, while Turnadge et al., (2018a) provided a refinement of water balance components and conceptual model.

In the following Section 3.4.1, we will present the components that constitute the groundwater balance of the Wildman River Area. The different components will be computed based on parameter values described in Section 3.4.2 and on the alternative conceptual models described in Section 3.4.3. The parameter values and conceptualizations used here are based on abovementioned investigations.

3.4.1 Water balance components

The simplified groundwater balance of the Wildman River Area is written as:

$$0 = Q_R - Q_L - Q_B + \Delta S + \delta \qquad (3.11)$$

where Q_R is the net recharge to the water table from rainfall accounting for the losses due to plant transpiration or direct soil evaporation (Doble and Crosbie, 2017). Net recharge is calculated based on a recharge area (A_R) and a recharge rate (R_R):

$$Q_R = A_R \cdot R_R \qquad (3.12)$$

Q_L is the lateral groundwater outflow from the aquifers to adjacent areas. The lateral outflow for the Wildman River Area is calculated through cross-sections based on Darcy's law. Darcy's law depends on the transmissivity of the aquifer (T_l), the width of the aquifer (W_l) and the hydraulic gradient perpendicular to the cross-section (Δh_l):

$$Q_L = T_l \cdot W_l \cdot \Delta h_l \qquad (3.13)$$

The lateral discharge consists of up to four subcomponents; lateral discharge across a northern boundary and a north-eastern model domain boundary, through the Mz/Cz sand aquifer and finally through the Koolpinyah Dolostone aquifer (Figure 3.3). The remaining boundaries in the model domain in south and northwest is bounded by impermeable (an order of magnitude less transmissivity) bedrock.

Q_b is the baseflow from the aquifer to streams and lagoons. In every time step (t), total streamflow (y_t) consists of overland flow that reaches the stream quickly, hence named quick flow (f_t), and baseflow (Q_B) that originates from groundwater discharge. The fraction of the streamflow that makes up the baseflow component is described by the baseflow index(β) (Eckhardt, 2008):

$$Q_{B,\text{streams}} = y_t - f_t = \beta \cdot y_t \tag{3.14}$$

The baseflow to streams consist of up to five subcomponents; Jimmy’s Creek, Opium Creek, Swim Creek, Cattle Creek, Ben Bunga Creek (Figure 3.3). The groundwater discharge to lagoons can be calculated based on seepage rate (R_L) and lagoon area (A_L):

$$Q_{B,\text{lagoons}} = A_L \cdot R_L \tag{3.15}$$

The same seepage rate is assumed for all lagoons in the area.

The final water balance component relates to groundwater storage, ΔS. Based on the observation the groundwater level returns to the similar level after each year (Turnadge et al., 2018a), the annual storage is assumed to be around 0 m^3.

For the Wildman River area, transient data for each of the water balance components is scarce, therefore a steady-state water balance is considered. However, the Wildman River Area is quite dynamic as the aquifers are filled up in the wet season by the high rainfall events and starts emptying again in the dry season when only very limited rainfall occurs. Given the large difference between wet and dry season, we will set up a steady-state water balance for both the end of the dry season (1. December) and the end of the wet season (1. April). The wet season and dry season model will be combined into an annual model, so that:

$$\delta = \left(Q_{R,\text{wet}} + Q_{R,\text{dry}}\right) - \left(Q_{L,\text{wet}} + Q_{L,\text{dry}}\right) - \left(Q_{B,\text{wet}} + Q_{B,\text{dry}}\right) + (\Delta S_{\text{wet}} + \Delta S_{\text{dry}}) \tag{3.16}$$

3.4.2 Water balance parameters

An overview of the parameter values used in the stochastic water balance model is shown in Table 3.2. A discussion of the derivation of the prior parameter distributions is found in Appendix B.1- B.4. Uniform distributions are used for all parameters defined by the minimum and maximum value, so that all parameter values in the ranges are equally likely. All the defined rates are used as spatial averages over the whole area.

Table 3.2. Parameter values describing the water balance components in the Wildman River Area. The parameters are described with a uniform distribution between the minimum and the maximum value in the dry and in the wet season. Parameters that describe areas, width and transmissivities are constant over the year, and therefore only described by one set of a minimum and a maximum value. The term sand refers to Mz/Cz sand, while the term Dolostone refer to the Koolpinyah Dolostone.

Component	Parameter	Dry Min	Dry Max	Wet Min	Wet Max	Unit
Net Recharge	Rate	0	0	32	178	mm/y
	Area	350	400	350	400	km^2
Lateral discharge	Transmissivity Dolostone	109	2630	109	2630	m^2/d
	Transmissivity Sand	163	1920	163	1920	m^2/d
	Gradient North	0.0003	0.0009	0.0004	0.0012	-
	Gradient Northeast	0.0002	0.002	0.0004	0.004	-
	Width Dolostone North	3000	10000	3000	10000	m
	Width Dolostone Northeast	1000	7000	1000	7000	m
	Width Sand North	1000	13000	1000	13000	m
	Width Sand Northeast	1000	7000	1000	7000	m
Lagoons	Area	2.9	3.2	2.9	3.2	km^2
	Rate	0.5	2	0.5	2	mm/d
Streams/ springs	Baseflow Jimmy’s Creek	0.06	0.09	0.2	0.3	m^3/d
	Baseflow Opium Creek	0.05	0.07	0.2	0.3	m^3/d
	Baseflow Swim Creek	0.03	0.1	0.7	2.1	m^3/d
	Discharge Cattle Creek	0.001	0.005	0.005	0.03	m^3/d
	Discharge Ben Bunga Creek	0.001	0.005	0.005	0.03	m^3/d
	Base Flow Index	0.22	0.81	0.22	0.82	-
Annual storage		0	0	0	0	-

3.4.3 Alternative conceptual models

Even though many investigations (Tickell and Zaar, 2017; Turnadge et al., 2018a, 2018c) have been carried out in the Wildman River Area recently, there are still several open conceptual questions. In this paper we focus on the conceptual issues that have a direct influence on the annual water budget of the catchment. In the following discussion, five uncertain water balance components will be identified, and alternative hypotheses will be defined that demonstrate the conceptual issues; in a subsequent step their influence on the water balance will be quantified. The definition of hypotheses will be based on the method specified in Section 3.3.1. An overview of the conceptual model and conceptual uncertainties in the Wildman River Area is shown in Figure 3.3.

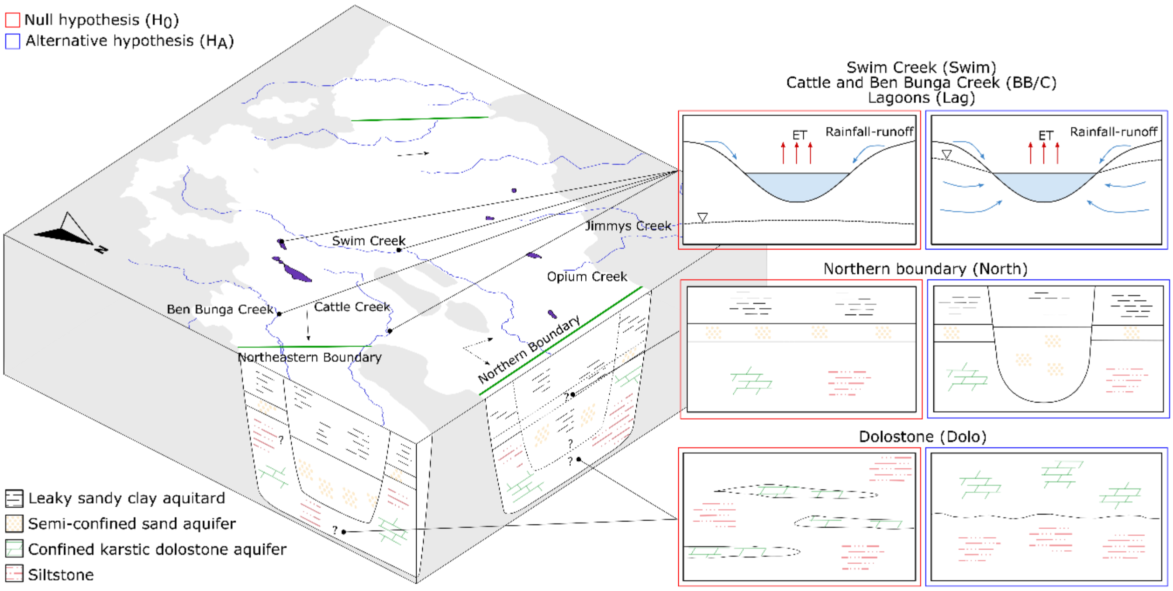
□ Null hypothesis (H_0)
□ Alternative hypothesis (H_A)
Swim Creek (Swim)
Cattle and Ben Bunga Creek (BB/C)
Lagoons (Lag)
ET
Rainfall-runoff
ET
Rainfall-runoff
Swim Creek
Jimmys Creek
Opium Creek
Ben Bunga Creek
Cattle Creek
Northeastern Boundary
Northern Boundary
Northern boundary (North)
Dolostone (Dolo)
?
N
□ Leaky sandy clay aquitard
□ Semi-confined sand aquifer
□ Confined karstic dolostone aquifer
□ Siltstone

Figure 3.3. Conceptually *uncertain components in Wildman River Area and hypotheses developed to characterize the conceptual uncertainty.*

Given the spatially sparse groundwater level observations in the study area, the groundwater flow around the northern boundary of the system can be interpreted in different ways. In one instance (Tickell and Zaar, 2017), groundwater was considered to flow north across the boundary along a northern palaeovalley (i.e. out of the domain) contributing to lateral discharge in the water balance. However, observations may also indicate a northeastwards groundwater flow, in which case groundwater will flow along the northern boundary rather than across, resulting in limited lateral discharge across the boundary. Two alternative hypotheses regarding the northern boundary component of lateral discharge are defined:

- H_0: A northern palaeovalley does not exist and the groundwater flows along (i.e. parallel to) the northern boundary of the system and therefore no lateral discharge into or out of this area occurs.
- H_A: A northern palaeovalley exists and the groundwater flows across the northern boundary and therefore contributes to the total lateral discharge out of the model domain.

Based on borehole observations and the location of what is inferred to be sinkholes developed above the Dolostone, Tickell and Zaar (2017) have interpreted the extent of the Koolpinyah Dolostone as a relatively continuous aquifer. However, the fact that the basement geology including the Koolpinyah Dolostone is folded tightly indicates the plausibility that the Dolostone aquifer consist of more or less structurally isolated aquifers. In this case only the Mz/Cz sand aquifer would contribute to lateral discharge while the Dolostone aquifer would not. This leads to the following hypotheses regarding the Dolostone component of lateral discharge:

- H_0: The Koolpinyah Dolostone is a compartmentalized aquifer and therefore its contribution to lateral discharge is unimportant.
- H_A: The Koolpinyah Dolostone is a continuous aquifer and contributes significantly to lateral discharge.

Ben Bunga and Cattle Creek are ephemeral streams that drain northeastwards towards Kakadu National Park. They cease to flow in the early dry season but maintain several isolated permanent pools that are hypothesized to receive groundwater flow through diffuse streambed discharge (Tickell and Zaar, 2017). However, their ephemeral nature makes this a questionable assumption. It is considered very unlikely that only one of these creeks would receive groundwater and the other not; therefore, the hypotheses for both creeks are combined into a single one:

- H_0: Ben Bunga and Cattle Creek are a rainfall-runoff feature, disconnected from the groundwater system.
- H_A: The streamflow in Ben Bunga and Cattle Creek originates from both groundwater discharge and rainfall-runoff.

Swim Creek is an ephemeral stream that drains the central region of the Wildman River Area and flows northwards. It has ceased to flow at the end of the dry season in 15 out of 29 years of the recorded stream flow. The baseflow index was previously estimated to be around 50 % (Tickell and Zaar, 2017). A different conclusion was obtained by comparing the streamflow record from Swim Creek to that from Opium Creek (Turnadge et al., 2018a): streamflow for the former is generally an order of magnitude larger than that for the latter. This was attributed to the much larger surface water catchment of Swim Creek, which led to the assumption that Swim Creek is primarily fed by rainfall-runoff. The hypotheses regarding Swim Creek are defined as follows:

- H_0: Swim Creek is a rainfall-runoff feature, disconnected from the groundwater system.
- H_A: Streamflow in Swim Creek originated from groundwater as well as rainfall-runoff.

A large number of shallow depressions exist in the Wildman River area that are interpreted to be sinkholes formed on top of the Koolpinyah Dolostone (Tickell and Zaar, 2017). Some of these shallow depressions serve as permanent water features, referred to as lagoons, which has led Turnadge et al. (2018a) to hypothesize that they are groundwater discharge features. However, using a mass balance approach, groundwater discharge to the lagoon was found to only occur during wetter than average climate conditions (Tickell and Zaar, 2017). Until date only the largest of the Twin Sisters lagoons has been subject to investigations, involving the comparison of lake stage recession to evaporation rate (Graham, 1985). Based on one year of observations, it was estimated the lagoon was a flow-through feature. In yet another study, analysis of the noble gas tracer Rn-222 in surface water samples collected from the lagoon, also did not yield conclusive results in regards to whether or not groundwater inflow occurs (Turnadge et al., 2018a). These ambiguous findings around potential groundwater contribution to Twin Sisters Lagoon makes this an important conceptual uncertainty. The other permanent lagoons in the area (Number One Billabong, Lake Lucy and Mistake Billabong) are assumed to behave in the same way as the Twin Sisters lagoons. The hypotheses regarding permanent lagoons are defined as follows:

- H_0: The permanent lagoons are rainfall-runoff features, disconnected from the groundwater system.
- H_A: The permanent lagoons are, at least in part, groundwater discharge features.

In the above discussion, a total of five uncertain water balance components have been identified, while two alternative models have been defined for each uncertain component. Using the factorial design approach described in Section 3.3.1 thus gives 2^5 or 32 individual models that will be quantitatively evaluated.

We assign a uniform prior probability to the 32 alternative combinations, i.e. all models are considered equally likely. By using uniform priors, we expect that the evaluation of internal model consistency expressed as the likelihood function (Section 3.3.4) dominates the resulting posterior distribution. Alternatively, the prior could have been based on an expert elicitation process, as in (e.g. Meyer et al., 2007; Ye et al., 2008b), to be able to further differentiate between the models. However, the scope of the paper is to evaluate whether there is enough information in the water balance offer insight into the conceptualizations. Expert elicitation of the prior probabilities is therefore beyond the scope of this paper.

3.5 RESULTS

In this section, we apply the Bayesian hypothesis testing framework to the stochastic water balance in the case study. On a 2.4 GHz computer with 8 GB RAM every 10.000 realisation takes ~ 1s with an acceptance rate of the water balance realizations of ~ 27 %. A graphical description of the setup of the model is seen in Figure 3.1. The illustration shows that the forward model, i.e. the groundwater balance problem, is linear and additive, and the setup allows for vectorization and parallelization, which allows for evaluation of a large set of model versions. All implementations, calculations and sampling is performed in the Python 3.6 computing environment with the software stack of NumPy (Oliphant, 2006) while the figures are prepared with Matplotlib (Hunter, 2007).

3.5.1 Posterior probabilities of hypotheses based on assumed error

The simple and conditional probabilities for the alternative hypotheses (HA) are shown in Figure 3.4. The simple probabilities shown in Figure 3.4a is the marginal probability of a subset of 16 out of the 32 different models (H0 vs HA). The conditional probability shown in Figure 3.4b is the marginal probability of a subset of 8 out of the 32 different models that meets the conditions described in the parentheses. The corresponding probabilities for the null hypotheses (H0) (only shown for Figure 3.4a) can be obtained as 1 subtracted the probabilities for alternative hypothesis. By applying Bayesian hypothesis testing, we have implicitly assumed that a quasi-true model can be identified from the alternative model ensemble (Höge et al., 2019). Given more data, the probabilities for H_A or H_0 would therefore further approach either 1 or 0.

The simple probabilities (Figure 3.4a) show that a clear preference for the alternative hypothesis is supported only for the conceptually uncertain Swim Creek with a value of 39 % and 61 % for the null and alternative hypothesis, respectively. The support can, however, still only be described as “Inconclusive” (Section 3.3.3), based on a Bayes factor of less than 3 (63.5 %, Table 3.1). For the rest of the conceptual uncertain components, the change between the prior probability of 0.5 and the posterior probability is even smaller. This indicates that the information content in the closure of the water balance is too small to sufficiently differentiate between models for valid model rejection to occur. A balanced water budget can result from sufficiently accounting for all water balance components, but can also arrive from globally balanced errors (Dassargues, 2018). The results therefore suggest that without other constraints, all suggested conceptual models are valid because they are either true or can sufficiently balance errors in the conceptualization globally. A possible strategy to increase model discrimination is to further constrain parameter priors by collection of more data (Pham and Tsai, 2016). With further constraints on the parameter priors, i.e. reduction of parameter ranges in Table 3.2, the ability of the different models to balance errors globally reduces as the magnitude of the water balance components will vary less. Another strategy to increase the ability to differentiate between models is to apply informed priors, but as already stated in Section 3.4.3, this is beyond the scope of this paper.

The conditional probabilities (P(X|Y)) (Figure 3.4b) describe the probability of a hypothesis X given the assumption that another hypothesis Y is true. It thereby offers a preview of how the probabilities would change for hypothesis X if we found hypothesis Y to be true, e.g. by collecting additional data.

Provided the alternative hypothesis of any uncertain component, but Swim Creek, the conditional probabilities are higher in case an alternative hypothesis ($P(X\ H_A|Y\ H_A)$) rather than a null-hypothesis ($P(X\ H_A|Y\ H_0)$) is used as the other given event (i.e. other uncertain component). This indicates a general preference for applying more components to balance the water budget. That is, if none of the additional uncertain discharge components (alternative hypotheses) are applied, there is a surplus in the water balance. However, when Swim Creek model component is involved, a trade-off with the remaining alternative hypotheses can be observed. This is especially true for the conditional probability of the alternative hypothesis for Swim Creek given the Dolostone (P(Swim H_A|Dolo H_0)) and the northern boundary (P(Swim H_A|North H_0)) hypotheses, where the support becomes "substantial" (both 0.65), when the null-hypotheses are applied. This indicate that the input to the water balance is not large enough to account for both the alternative hypotheses for Swim Creek and the Dolostone or Northern boundary, that all include extra discharge terms. In conclusion, the largest change in conditional probability can be observed for when Swim Creek is involved and future field work should therefore aim at resolving this conceptual uncertainty first.

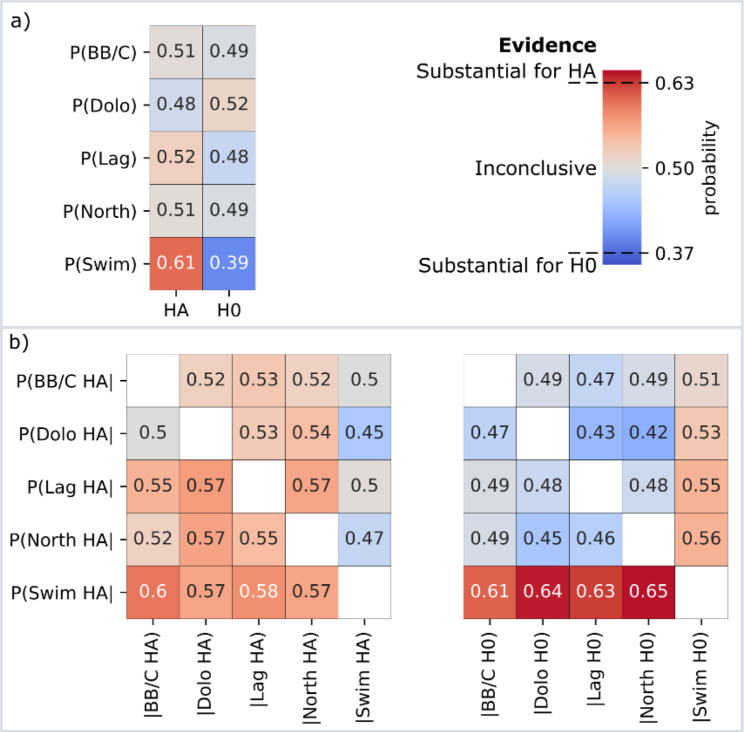

Figure 3.4. Simple (a) and conditional (b) posterior probabilities of hypotheses concerning uncertain water balance components. (b) should be read: probability of row, given column, e.g. P(Dolo HA|North HA). In (b) the inverse probability (not shown here) is P(X HO|Y HA) = 1 - P(X HA|Y HA) and P(X HO|Y HO) = 1 - P(X HA|Y HO), respectively. All started with uniform probability (0.5). In Figure 3.2, names in this figure refers to: BB/C = Ben Bunga and Cattle Creek, Dolo = Dolostone, Lag = Lagoons, North = Northern boundary and Swim = Swim Creek.

3.5.2 Model Predictions

The prior and posterior predictions for recharge, lateral discharge and baseflow as obtained from the stochastic water balance calculations are shown in the top row of Figure 3.5. The obtained predictions represent a multi-model probability density function that takes account of all the conceptual models that seem plausible under the current state of knowledge as well as the parameter uncertainty. These results are compared with the deterministic estimates for the different water balance components in Tickell and Zaar (2017), who provided two independent water balance estimates, shown as vertical lines in Figure 3.5.

Compared to the deterministic solutions from Tickell and Zaar (2017), the posterior probability has not changed significantly. However, in our stochastic predictions, both parameter and conceptual uncertainty are accounted for, we now have a water balance of which the confidence limits are quantified.

The prior probability for baseflow is highly bimodal, caused by the trans-dimensional sampling between models that include extra baseflow components (alternative hypotheses for Ben Bunga and Cattle Creek, Lagoons and Swim Creek) and the models that exclude the extra baseflow component (corresponding null hypotheses). The prior probability for recharge and lateral discharge is however unimodal, suggesting that the conceptual uncertainty is of less importance than the parameter probability.

The posterior probability for recharge and baseflow is multimodal (Figure 3.5, top row). While the prior probability for the baseflow is already multimodal, the prior for recharge is not and the shape of the posterior is therefore caused by the conditioning to the closure of the water balance. Overall the range of the posterior of the recharge is shown to be reduced to a maximum of 60 GL/y, whereas the maximum for the prior distribution is 70 GL/y. However, the posterior probability of the lateral discharge and baseflow has not changed significantly, indicating low information in data (the balancing of the water budget), because a trade-off exist between the output components. Both lateral discharge and baseflow are loss terms in the water balance that will respond in a similar way when poorly defined parameter distributions are used, as is the case here.

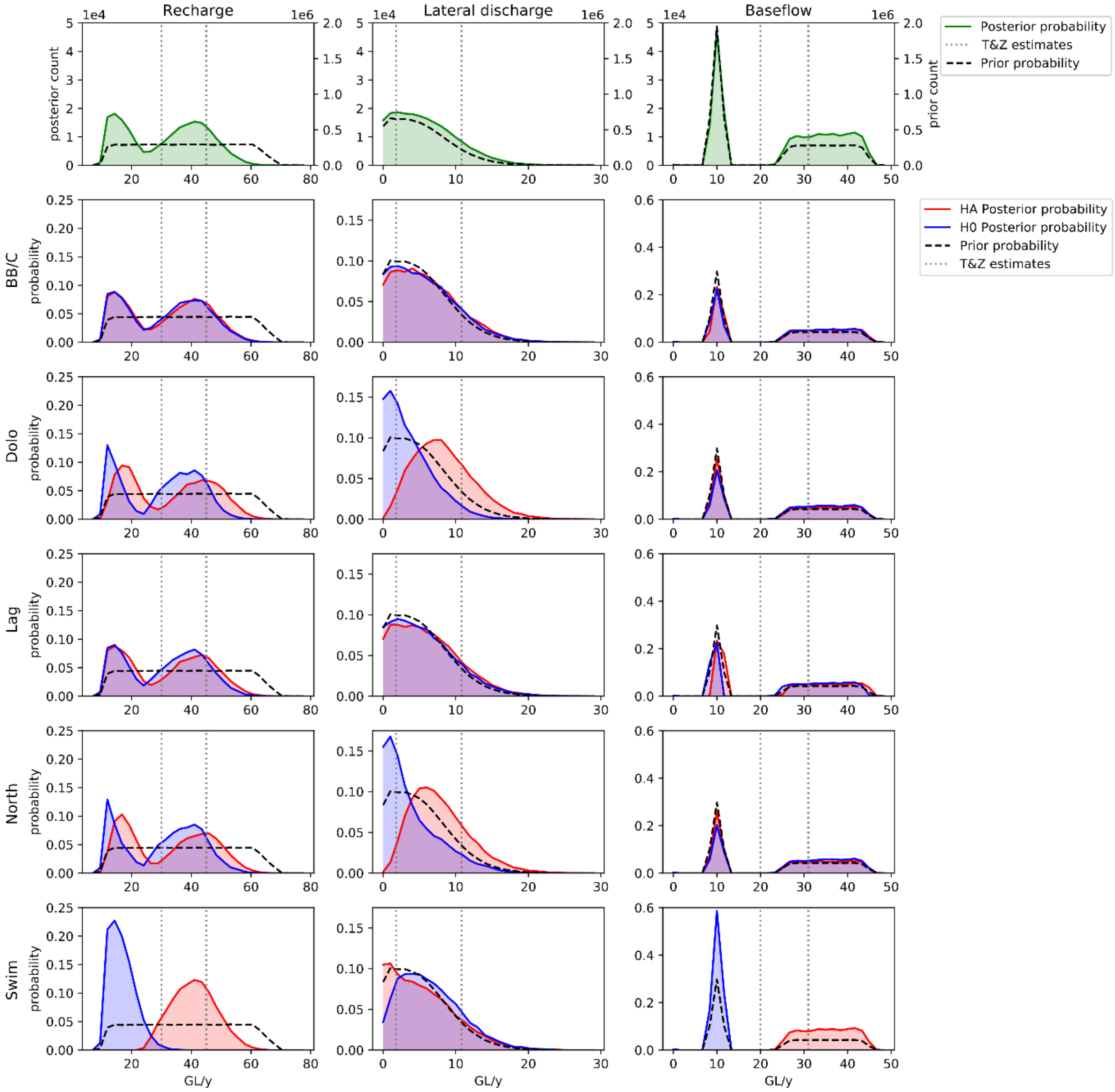

Figure 3.5. Prior and posterior probability of water balance components recharge, lateral discharge and baseflow. Top row shows the probabilities for all conceptual models, while in the remaining rows the posterior probabilities are subdivided into the null and the alternative hypothesis regarding the row. Individual estimates from (Tickell and Zaar, 2017) *of the three water balance components is shown as vertical lines. Note the difference in scale for the prior and posterior probabilities.*

The impact of different hypotheses on the simulated recharge, lateral discharge and baseflow is shown in Figure 3.5; each row represents posterior probability subdivided into the models that includes the null and the alternative hypotheses (16 models each) for one of the five uncertain components, as referenced in the row header. The within model variance is represented by either the red or blue probability distribution depending on the chosen model, while the between model variance is represented by the difference between the red and the blue probability distribution.

In three out of five cases, we can observe an increase in the amount of recharge that can be supported in the model when an alternative hypothesis (blue) is applied, as all alternative hypotheses add an extra discharge component to the water balance. In only two cases recharge does not increase: i.e., for the conceptual uncertainty regarding Ben Bunga and Cattle Creek (row BB/C) and the Lagoons (row Lag). While these components directly impact the baseflow prediction, they are shown to have very little impact on recharge and lateral discharge. The conceptual uncertainty regarding the Dolostone (row Dolo) and the northern lateral boundary (row North), both directly impacting the lateral discharge prediction, is however shown to have a more pronounced indirect impact on recharge (i.e. the probability distribution shifts to higher values by about 10 GL/y). The largest impact on the overall predictions is, however caused by the conceptual uncertainty regarding Swim Creek (row Swim). The trans-dimensional sampling between including and excluding the discharge component from Swim Creek hypothesis directly impacts baseflow, which becomes bimodal, and thereby indirectly affects the prediction of recharge which also becomes bimodal.

Again, it is shown that the reduction of uncertainty would be greatest if we were able to resolve the Swim Creek conceptual uncertainty e.g. by additional field work. The additional field work could target Swim Creek directly, but it could also aim at reducing the aleatory uncertainty of net recharge component. Figure 3.5 shows that if the net recharge to the area is more than 40 GL/y the alternative hypothesis for Swim Creek is true. However, if the net recharge to the area is less than 20 GL/y the null hypothesis for Swim Creek would be true.

3.6 DISCUSSION

The results presented in Section 3.5 are consistent with what is presented in other groundwater modelling studies (e.g. Rojas et al., 2008, 2010c, 2010a; Schöniger et al., 2015a; Troldborg et al., 2010; Zeng et al., 2015), in that we have obtained posterior model probabilities and a probability distribution of predictions including both parameter and conceptual uncertainty. However, the methods with which these model results have been obtained, differ. The advantages and disadvantages of our approach compared to the above-mentioned studies is discussed in the following.

We applied a factorial design approach, where all possible combinations of hypotheses are tested, which enables the attribution of differences in model performance, directly to specific conceptually uncertain components. The factorial design approach can isolate the causes affecting the model predictions. However, as every new conceptually uncertain component doubles the number of possible models (assuming two hypotheses are defined for each uncertain component), the problem can quickly become time consuming. This practical barrier is referred to as the fallacy of factorial design in (Betini et al., 2017).

To avoid making the problem numerically intractable, we have applied a very simple additive and linear model setup (rather than a 3D numerical groundwater flow or transport model), which enables us to run millions of models in the matter of seconds. The studies we compare our results to run between 4,000 and 300,000 realizations. By being able to run more realizations, we ensure a more stable result and that there is no practical limitation to how many different conceptual models can be evaluated. The simplicity of the model setup allows us to gain insight, without undue amount of time, which can be brought forward into subsequent more complex modelling.

The disadvantage of our simple approach is that the data the models is tested against is limited. In the applied setup the only data used to evaluate the models against, is the assumption that the water budget is balanced. All other data in the case study was used to set up the priors for the parameters describing the water balance. In contrast, the above-mentioned studies included testing data such as hydraulic head, contaminant concentrations and pumping tests. The limited available evaluation data in the suggested approach means that we will not be able to discriminate between models to the same extent as the above-mentioned studies. To improve the discriminatory power of this modelling approach more data should be reserved for model evaluation.

The model cannot underpin environmental management but is an initial screening tool built to allow the modeler to gain insight into the system functioning, identify important sources of uncertainty and prioritize research efforts. In a stepwise approach to groundwater modelling (discussed in Section 3.2), the suggested approach would constitute one of the initial steps after the hydrogeological characterization before moving towards testing in a more complex mathematical model. This model testing step would then inform the succeeding steps ensuring a transparent workflow.

3.7 CONCLUSION

We presented an approach to model-based Bayesian hypothesis testing in a simple additive groundwater balance model, which involves optimization of a model in function of both parameter values and conceptual model. The proposed systematic conceptual model development method allows for directly attributing the differences in performance of alternative models to individual uncertain components in the conceptual model.

The method was demonstrated on a water balance model for Wildman River Area. Five conceptually uncertain components resulted in 32 individual conceptual models and millions of realizations with all conceptualizations being conditioned to the closure of the water balance. The following can be concluded from the case study:

- More confidence has been gained in the water balance compared to the deterministic solution. Probabilistic distribution of predictions take account of all the conceptual models that seem plausible under the current state of knowledge as well as the parameter uncertainty.
- The understanding of the system functioning has increased. None of the conceptual models can be ruled out, but we have a better idea of how important they are to the water balance predictions and how they impact parameter ranges.
- The fieldwork going forward can now be prioritized in terms of the impact the different components have shown on the water balance predictions.

Testing alternative conceptual models is recognized to increase transparency, help prioritize research effort and help uncover potential conceptual surprises. The overall conclusion of this study is that testing alternative conceptual models does not have to be a time-consuming task, but can be done in relatively simple models, e.g. as here, in a water balance model.

Chapter 4: A Systematic Approach to Hydrogeological Conceptual Model Testing

Trine Enemark, Luk JM Peeters, Dirk Mallants, Okke Batelaan, Brady A. Flinchum

Submitted to *Water Resources Research*[3]

[3] Enemark, T., Peeters, L. J. M., Mallants, D., Batelaan, O., & Flinchum, B. A. (submitted). A systematic approach to hydrogeological conceptual model testing combining remote sensing and geophysical data.

4.1 ABSTRACT

Conceptual uncertainty is considered one of the major sources of uncertainty in groundwater flow modelling. Hypothesis testing is essential to increase system understanding by analysing and refuting alternative conceptual models. We present a systematic approach to conceptual model testing aimed at finding an ensemble of conceptual understandings consistent with prior knowledge and observational data. This differs from the traditional approach of tuning the parameters of a single conceptual model to conform with the data through inversion. We apply this approach to a simplified hydrogeological characterisation of the Wildman River Area (Northern Territory, Australia) and evaluate the connectivity of sinkhole-type depressions to groundwater. Alternative models are developed representing the process structure (i.e. different fluxes representing interactions between surface water and groundwater) and physical structure (i.e. different lithologies underlying the depressions) of the conceptual model of the depressions. Remote sensing data are used to test the process structure, while geophysical data are used to test the physical structure. Both data types are used to remove inconsistent models from an ensemble of 16 models and to update the probability of the remaining alternative conceptual models. Three out of five depressions that are used as a test case are conditionally confirmed to act as groundwater recharge features, while for the last two depressions, the data is inconclusive. Although the framework is not directly prediction oriented, the testing of plausible conceptual models will ultimately lead to increased confidence of any groundwater model based on accepted posterior conceptualisations.

4.2 INTRODUCTION

The conceptual understanding of groundwater systems is widely recognized as a major source of uncertainty in hydrogeological model predictions (e.g. Refsgaard et al., 2012). Traditionally a single conceptual model forms the basis for the model predictions. However, the available data on the groundwater system often support more than one conceptualisation. Rejecting all but one plausible conceptual model by omission presents serious issues for the hydrogeological modelling workflow. The overall predictive uncertainty is at best underestimated; moreover, the model prediction may be biased due to the possible choice of an invalid model (Neuman, 2003). Further, the very choice of the most representative conceptual understanding of a system may be an ad hoc task (Clark et al., 2011) which presents a reproducibility issue for the groundwater modelling workflow.

In this case, conceptual model testing presents a promising tool to increase transparency, reproducibility, and to integrate an automation of an expert's thought in the modelling workflow (Enemark et al., 2019a). In model testing alternative understandings are proposed and independent data are used to attempt to refute the alternative concepts. Model testing allows for a transparent account of model choices, rejection of invalid conceptual models and unveiling of conceptual "surprises" (Ferré, 2017). By reporting alternative conceptual models and applying model testing, the confidence in the model predictions increases and the risk of potential bias decreases (Hassan, 2004).

In literature the objective of model testing is often to reduce the number of plausible alternative conceptual models. Approaches to choose between alternative conceptual models was classified into three broad categories by Carrera et al. (1993): 1) Comparative analysis of predicted and observed values (Pirot et al., 2015; Zeng et al., 2015), 2) assessment of calibrated parameter values to observed values (Engelhardt et al., 2014; Poeter and Anderson, 2005), and 3) "identification criteria" (model selection criteria) which are often based on maximum likelihood and may also consider the principle of parsimony (Schöniger et al., 2014).

The currently applied testing approaches in hydrogeology suffer from several challenges. One of the challenges in conceptual model testing is to prioritize independent data for testing rather than using all available data for development of models. Testing data must be independent from model development data in order to avoid circular reasoning, overconfidence in the conceptual models and under-sampling of the model space (Chatfield, 1995; Kerr, 1998).

Without planning on prioritizing independent data for model testing, any data previously used for model development, might have little power to test the model (Rojas et al., 2010c). The power of the model testing data relates to the type, amount and uncertainty of that data. The consequence of testing alternative model conceptualizations with data that have limited information content is that discrimination among alternative models cannot be made (Enemark et al., 2019b; Seifert et al., 2008) and, in a Bayesian context, that the prior probability has a large influence on the posterior probability (Rojas et al., 2009).

Finding existing data that is independent from that used for model development is often a challenge, and the best approach to ensure independence is probably to collect an entirely new data set (sometimes referred to as a post audit) (Anderson and Woessner, 1992).

Another challenge is determining when models can be considered rejected. In hydrogeology, models are frequently rejected after the model probability has been updated. Rejection is typically based on a very low probability (Hermans et al., 2015; Park et al., 2013) or based on a threshold value for the Bayes Factor (Brunetti et al., 2017). These forms of model rejection require that another model exists that outperforms the rejected model (Gelman and Shalizi, 2013); and it is therefore implicitly assumed that the range of alternative models is collectively exhaustive (Enemark et al., 2019a). Further, the best performing model becomes unfalsifiable, meaning that conceptual "surprises" cannot be uncovered even if the model that most adequately represents the real system has not been developed yet.

Hypothesis testing applying a Bayesian framework combined with falsification type approach is sometimes referred to as Popper-Bayes philosophy (Gelman and Shalizi, 2013; Linde et al., 2015b; Tarantola, 2006). The Popper-Bayes philosophy address the challenge of when to reject models. This hypothetico-deductive framework builds on the idea that "observations cannot produce models; they can only falsify models". The falsification type approach consists of checking the models against data and rejecting models that are inconsistent. The Bayesian framework on the other hand, offers a systematic approach to updating the prior beliefs about the adequacy of a model to posterior beliefs.

The framework applied in this paper is shown in Figure 4.1. After developing the alternative model(s) (step 1), independent data is collected (step 2) to ensure the testing of the models is indeed independent thus avoiding circular reasoning. Forward models can then be run (step 3) to provide input to model rejection (step 4) after which the probability will be updated (step 5).

This approach differs from tuning the parameters of a single conceptual model to conform with the data through inversion. Traditionally, the difference between a forward modelled response and the observed response is used to drive an inversion to find the model that best explains data. Usually the model with the lowest mismatch between observed and modelled response is considered the best model. In the methodology applied here, instead of relying on the inversions which are prone to spatial averaging and smoothing to interpret the conceptual model, we use the forward modelled data in a unique way to ask the question, how well do our forward model response by our conceptual model fit the observed data? The method aims at finding an ensemble of conceptual understandings that can be sufficiently explained by data, rather than finding the model that best fit data through a process of inversion.

Conceptual model testing with similar frameworks has been successfully applied in several hydrogeophysical case studies (Linde, 2014). Model testing in hydrogeology without use of geophysical data and with more than one data type is however still largely unexplored with limited guidance in international literature and without workflows that can be readily adopted by practitioners.

The objective of this study is to provide a generally applicable workflow for systematic conceptual model testing that (i) updates prior belief in conceptual models through Bayes Theorem, (ii) based on diverse types of data that (iii) have not been used in the development of the conceptual model.

We apply this approach to a hydrogeological characterisation of the Wildman River Area (Northern Territory, Australia). More specifically we evaluate the connectivity of sinkhole-type depressions in the area to groundwater. We update the belief in the hypotheses around the model structures first using remote sensing data and then using seismic refraction data, focussing on the process structure and physical structure of the conceptual model, respectively. The analysis is focused on five sinkholes that were characterized based on past data and for which geophysical and remote sensing data were collected for model testing.

4.3 METHODOLOGY

The systematic testing approach for exploratory analysis of conceptual models is presented in Figure 4.1. The workflow starts by developing alternative models (step 1) of prior conceptualisations representing different understandings of groundwater system functioning. These models are then compared to data independent of data used for model development (step 2) through running a forward model (step 3) based on the alternative understandings.

The correspondence between multiple model realisations and observations can then be used to reject the unsupported models from the prior model ensemble (step 4) and update the probability of the remaining alternative understandings (step 5). When all alternative models are rejected or when new data for model testing is available, the workflow can be repeated.

In the following sections, we provide an overview of the individual components in the model testing workflow (Figure 4.1). This systematic conceptual model testing workflow is subsequently demonstrated on a case study area in the Wildman River Area, Australia, where sinkhole-type depressions are a common feature of the landscape, potentially contributing to highly localised groundwater recharge.

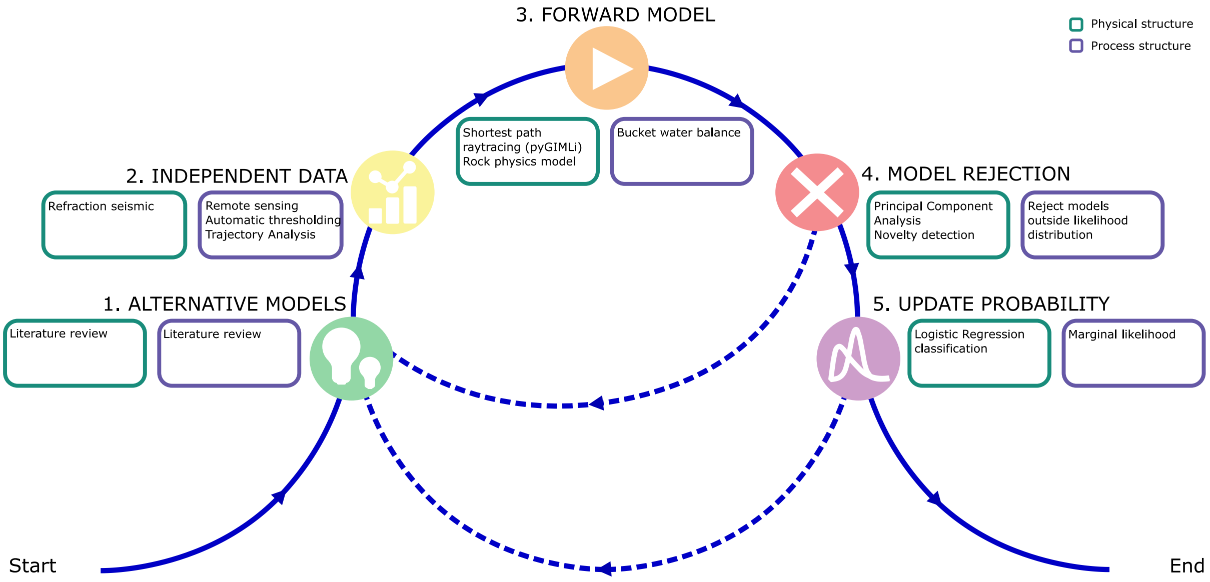
3. FORWARD MODEL
Physical structure
Process structure
Shortest path raytracing (pyGIMLi) Rock physics model
Bucket water balance
2. INDEPENDENT DATA
Refraction seismic
Remote sensing Automatic thresholding Trajectory Analysis
4. MODEL REJECTION
Principal Component Analysis Novelty detection
Reject models outside likelihood distribution
1. ALTERNATIVE MODELS
Literature review
Literature review
5. UPDATE PROBABILITY
Logistic Regression classification
Marginal likelihood
Start
End

Figure 4.1. Flowchart of systematic conceptual model testing for exploratory analysis. The objective of the approach is to start from prior uncertain conceptual models, test these with independent data and deliver posterior conceptual models that have a higher degree of confidence. The solid workflow lines are applied to the Wildman River area. In the application study the conceptual model is divided into a process structure and a physical structure and the methods applied in each step of the workflow are indicated in the boxes. The workflow has an iterative option (dashed lines) if/when new data is collected, or all alternative models are rejected.

4.3.1 Develop alternative models

Developing alternative conceptual models presents a natural first step of the systematic testing exercise (Figure 4.1). According to the definition by Gupta et al. (2012), a hydrogeological conceptual model is a summary of the current knowledge about the groundwater system. Any alternative understandings can therefore be based on a literature review of the site under investigation as well as general or subject-matter knowledge about groundwater system functioning (Chatfield, 1995; Clark et al., 2011). During the site literature review, one must identify the rationale underpinning each assumption to ascertain whether alternative understandings could be possible or not (Peeters, 2017).

Any hydrogeological conceptual model consists of a process structure and a physical structure (Carrera et al., 1993; Gupta et al., 2012). To fully characterise conceptual model uncertainty, both process and physical structure uncertainty must be considered. If the objective of model development is to explore conceptualisations, Caers (2018) and Guillaume et al. (2016) argue that the model development process should aim at making bold alternatives that maximize the difference between alternative models in order to gain the most insights from a potential model rejection.

In a Bayesian approach, the formulation of a prior belief in each alternative conceptual model is also a part of the model development step. The prior belief is most often defined as uninformed (e.g. Pham and Tsai, 2015) but can potentially be derived using expert opinion (e.g. Ye et al., 2008). Nearing et al. (2016) suggested that assigning probabilities should not be based on an individual component of a model but rather should be based on the whole model.

4.3.2 Independent data

The second step in the testing workflow involves identifying independent observation data (Figure 4.1). As discussed in the introduction, the requirement for the testing data is that the data are independent from the model development data (i.e. data used in step 1, Figure 4.1) and that the alternative conceptual models would lead to distinguishable observations of that data type.

Oreskes et al. (1994) argued that “the greater the number and diversity of confirming observations, the more probable it is that the conceptualization embodied in the [forward] model is not flawed”. When the objective of the testing exercise is to explore the model space, this holds especially true.

4.3.3 Forward models

The third step in the testing workflow (Figure 4.1) involves setting up forward models that represents the alternative models defined in the first step.

A forward model is a simplified mathematical description of the system that captures the main physical or process structures simulating the response of a given conceptual model and its parameter values. The response of the forward model is a synthetic dataset of the same datatype as the independent model testing data. In hydrogeology the forward model is most often a numerical groundwater model (Remson et al., 1980) or solute transport model (Thomsen et al., 2016; Troldborg et al., 2010), but the forward model can also be a geophysical forward model (Brunetti et al., 2017; Hermans et al., 2015) or a water balance model (Enemark et al., 2019b). The key with forward model is that it allows us to predict a measurable response of any given combination of parameters and model structures, that we can then compare to observations.

Simple forward models are often preferred over complex ones as simple models can offer insights into system understanding that can be obscured in more complex models (Haitjema, 2006; Hunt and Zheng, 2012). Also, “pragmatic constraints on time and budget limit the number of models that can be tested and fewer models are tested when they are more complex” (Refsgaard et al., 2012; Schwartz et al., 2017). On the other hand, testing more complex models enables a more holistic consideration of the system.

4.3.4 Reject models

Model rejection should ideally separate models that are both consistent with prior knowledge and with observations from models that are not. Model rejection (step 4, Figure 4.1) relies on extracting values or global patterns from the independent data (step 2) and comparing them to the same features of the forward model response (step 3) from the alternative conceptual models (step 1). What ’being consistent’ means must be defined before the forward model response is compared to the observations. Alternative conceptual models can be rejected through direct comparison of simulated and observed variables of state (Zeng et al., 2015), or through evaluation of model behaviour such as plume behaviour (Pirot et al., 2015) or groundwater flow patterns (Zyvoloski et al., 2003).

In case all alternative models are being be rejected, a conceptual “surprise” (Bredehoeft, 2005) has been uncovered and the model structure must be considered to be an unknown unknown. This means one should start over and develop new conceptualizations (as per step 1 in the workflow, Figure 4.1).

The remaining models are not validated *sensu strictu*, but their correspondence with the independent data supports their likelihood (Oreskes et al., 1994). The remaining models are therefore brought forward into the next model testing step where their prior probabilities are updated.

4.3.5 Update probabilities

The final step in the model testing workflow (Figure 4.1) is updating the prior probability of the remaining alternative model (after the model rejection step) to a posterior probability.

In a Bayesian approach a prior probability of each alternative model $p(m_k)$ is updated to a posterior probability $p(m_k|Y_k)$ using Bayes’ rule (Webb, 2017):

$$p(m_k|Y_k) = \frac{p(Y_k|m_k) \cdot p(m_k)}{p(Y_k)} \tag{4.1}$$

where the conceptual model element $m_k = [m_{pl}, m_{ps}]$ and the observed data $Y_k = [Y_{pl}, Y_{ps}]$ refer to either the physical structure (pl) or the process structure (ps), that are both part of the conceptual model. The physical structure describes the hydrostratigraphy and geometry of aquifers, while the process structure contains the internally and externally controlled boundary conditions such as heads and fluxes in a system (Gupta et al., 2012).

The prior probability of each model element $p(m_k)$ describes our belief in each model before any data is considered. The prior probability is therefore defined in the model development phase (step 1, Figure 4.1). If we can assume that the physical structure influences the process structure, then the priors can be defined as:

$$p(m_k) = p(m_{pl}) \cdot p(m_{ps}|m_{pl}) \tag{4.2}$$

The observed data types can be said to be independent of each other if a parameter in the physical structure testing data set does not depend on any parameter in the process structure testing data set and vice versa. If the datatypes are independent, the likelihood $p(Y_k|m_k)$ describing the probability of the observed data given the model is:

$$p(Y_k|m_k) = p(Y_{pl}|m_{pl}) \cdot p(Y_{ps}|m_{ps}) \tag{4.3}$$

The likelihood and thereby also the posterior probability rely on the relative performance of the individual model structures of the ensemble against data. As indicated in Figure 4.1, if new data exists/is collected the conceptual model probabilities can be updated again. The posterior conceptual model gives us an indication of how much confidence we have in different model components.

4.4 APPLICATION

In this section we apply the methodology to test if widespread, enclosed sinkhole-like depressions in the landscape of Wildman River Area, Australia, connected to groundwater. First, a brief description is provided about the depressions in Section 4.4.1. In Section 4.4.2-4.4.6, the individual steps in the suggested workflow (Figure 4.1), as they are applied to the case study area are discussed.

4.4.1 Depressions in the Wildman River area

The Wildman River area is located in the northern part of the Northern Territory, Australia. Most of precipitation (97 %) in the area occurs in the wet season from November to April (Turnadge et al., 2018a). The geology consist of sand and clay sediments of Mesozoic-Cenozoic origin ranging in thickness of less than 25 m to over 100 m and overly the basement geology of dolostone, siltstone and sandstone (Tickell and Zaar, 2017). The area has in recent years been subject to two major hydrogeological investigations (Tickell and Zaar, 2017; Turnadge et al., 2018a), but still several open questions remain regarding its hydrogeological conceptualization.

Approximately 100 km west of the field site, sinkholes are known to have developed on top of the dolomitic bedrock (Tickell, 2013). These sinkholes are generally rounded, broad, shallow depressions and often form closed water features that show phases of filling up with water in the wet season and drying out in the dry season (Schult and Welch, 2006).

In the Wildman River area the dolostone is known to be less continuous but similar depressions are found. A x-ray diffraction analysis found that the dolostone mineralogical composition is mainly dolomite with fractions of muscovite and quartz (Turnadge et al., 2018a) suggesting that dissolution and sinkhole formation may also have taken place. The locations of possible sinkhole features in the Wildman River Area have been mapped by and described further in Appendix C.1 (Figure 4.2) (Easey et al., 2016; Mueller et al., 2016; Tickell and Zaar, 2017).

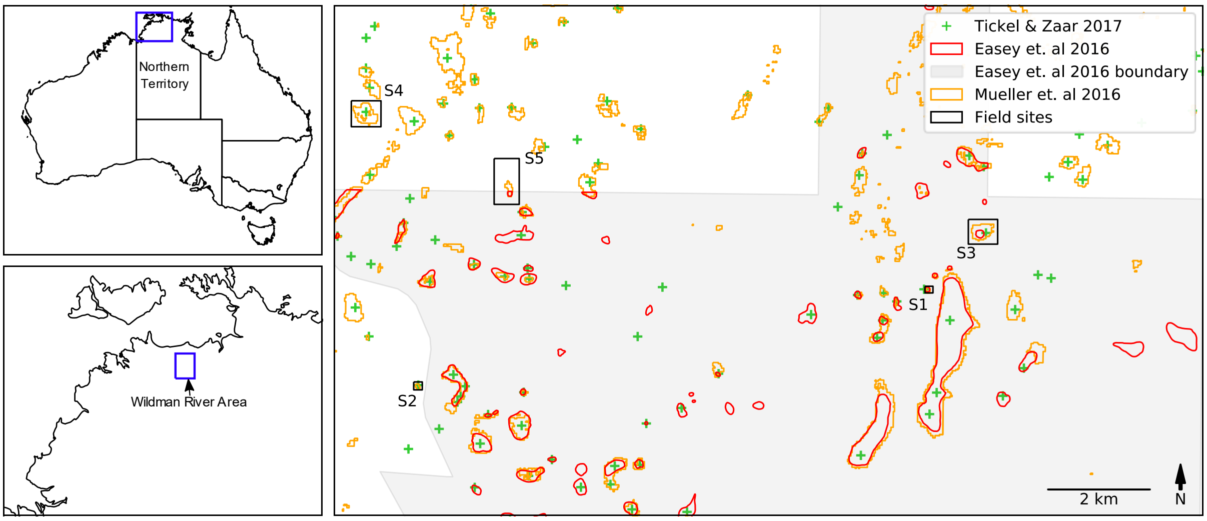

Figure 4.2. Mapped depressions in the Wildman River Area, Australia. The depressions that are part of the field investigation are marked with a black square and named S1 to S5.

At the end of the dry season (late October 2018), a fieldtrip was undertaken to investigate a subset of the mapped depressions in the Wildman River area. Five depressions (S1 to S5, Figure 4.2) were selected based on their accessibility and vicinity to existing boreholes and water level loggers. We prioritized the selection of sinkholes based on their difference in geometry and vegetation cover while making sure they covered different parts of the area. The collected data at each of the five depressions included a refraction seismic line, high water level marks, soil samples and topography. Further, PlanetScope satellite imagery has been collected over the area (Planet Team, 2017).

4.4.2 Conceptual model development

Previous investigations have hypothesized depressions in Wildman River area as sinkholes that may act as conduits for recharge (Graham, 1985; Turnadge et al., 2018a). This corresponds to observations that water levels in surficial aquifers respond within days of major rainfall events (Tickell and Zaar, 2017). Therefor, the null- and alternative hypotheses, respectively H_0 and H_A, tested in this paper are defined as follows:

- H_0: Depressions are connected to the groundwater.
- H_A: Depressions are not connected to the groundwater.

The hypothesis is that the depression are sinkholes, which in literature often is presented as groundwater flow conduits (Jardani et al., 2007; Kruse et al., 2006). Whether these sinkhole-like depressions act as conduits of recharge is an important conceptual question to resolve as it creates a potential for any contamination to rapidly reach the water supply.

The conceptual understanding is further refined in understandings of the process structure and physical structure (Figure 4.3). The physical structure accounts for different lithologies underlying the depressions (i.e. sand or clay), while process structure accounts for different interactions (i.e., fluxes) between the surface water and the groundwater. Process structure and physical structure are discussed in the subsequent sections.

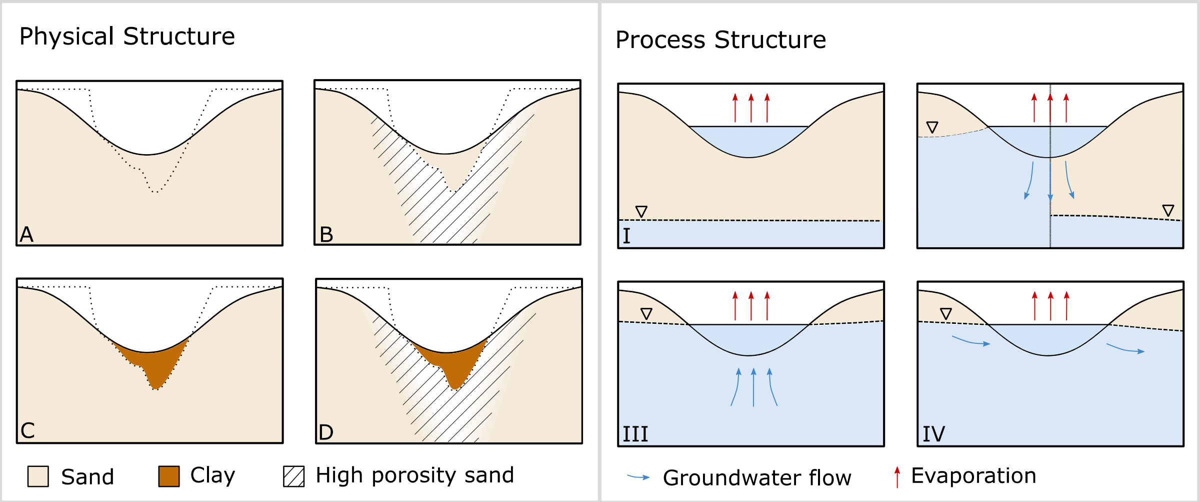

Figure 4.3. Alternative conceptual understandings of the physical and process structure of depressions in the Wildman River Area. The alternative understandings of the physical structure concern the presence (C, D) or absence (A, B) of a clay layer and the presence (B, D) or absence (A, C) of a high porosity zone in the middle of the depression. The dashed line in the physical structures represents an initial topography of the depression. The alternative understandings of the process structure consider whether the depression interacts (II, III, IV) or does not interact with the groundwater (I). Groundwater interactions tested are groundwater recharge (II), groundwater discharge (III) and whether the depressions are flow-through features with both groundwater discharge and recharge (IV). In model II we do not attempt to differentiate whether the groundwater recharge feature is directly connected or disconnected to the groundwater.

The alternative process and physical structures of the conceptual model of the depressions are shown in Figure 4.3. As stated in Figure 4.1, the alternative model structures are based on a literature review. A more detailed description of how these alternative structures have been obtained can be found in the Appendix C.2. A summary of process and physical structures is discussed below.

The process structures are based on a basic understanding of groundwater-surface water interactions of the depressions (Lloyd, 1999; Winter et al., 1998). The water balance of the depression is influenced by:

- I: Evaporation only, there is no exchange between surface and groundwater.
- II: Evaporation and groundwater recharge, where the groundwater flux occurs under either losing-connected or losing-disconnected conditions.
- III: Evaporation and groundwater discharge.
- IV: Evaporation, groundwater recharge and groundwater discharge.

The physical structures are based on an understanding of the geomorphological development of depressions (Graham, 1985; Schult and Welch, 2006; Tickell, 2013; Turnadge et al., 2018a):

- A: Homogeneous sand subsurface without vertical stratification.
- B: High porosity/permeability zone promoting water infiltration.
- C: Homogenous sand with vertical stratification through a sealing clay layer.
- D: High porosity/permeability zone overlain by a sealing clay layer.

The conceptual models are defined to have a maximum depth of 30 m because the depth of investigation of the refraction seismic data is limited to 30 m. Boreholes from the area show that the Dolostone is observed at depths between 40 and 100 m (Tickell and Zaar, 2017) and the model structures does therefore not include Dolostone. Note only the location and depth of investigation of the seismic survey was used to guide the design of the physical model structures, not the obtained seismic data values.

4.4.2.1 Prior probabilities

Current knowledge does not suggest any of the proposed model structures are more probable than others. However, the model structures are combined to (4 x 4 =) 16 different conceptual models and as the processes are mediated by the physical structure (Gupta et al., 2012), we assume that the specified physical structure influences the probability of the specified process structure. The groundwater exchange in the process structures (II, III, IV) combined with a sealing clay layer in the physical structures (C, D) is thereby assumed only half as probable as other combinations of the structures. This expresses our understanding of the physics of groundwater flow and the hydraulic barrier effect of clay.

Assuming a uniform probability otherwise and that the probabilities add up to one, models including a sealing clay layer combined with any groundwater exchange are assigned a probability of 0.0385, while all other contending models are assigned a probability of 0.0769 (Table 4.1).

Note that the prior probability assignment is subjective and the effect of these choices on the posterior probability could be evaluated by exploring the result of other prior probabilities. This is however outside the scope of this study.

It should also be noted that by assuming probabilities add up to one, we are implicitly assuming the range of models is collectively exhaustive, which is probably not the case. However, any remaining conceptual models are unknown unknowns.

Table 4.1. Prior probabilities of conceptual models consisting of a process structure and a physical structure.

Prior probability		Process Structure				
		I	II	III	IV	Total
Physical Structure	A	0.0769	0.0769	0.0769	0.0769	0.308
	B	0.0769	0.0769	0.0769	0.0769	0.308
	C	0.0769	0.0385	0.0385	0.0385	0.192
	D	0.0769	0.0385	0.0385	0.0385	0.192
	Total	0.308	0.231	0.231	0.231	1

4.4.3 Testing data

As indicated in Figure 4.1, the independent testing data for the process structure is remote sensing data, while refraction seismic data is used for testing the physical structure. All other collected data including soil samples, topography and water levels are used to parameterize the prior (Appendix C.5). A short overview of the independent data is given in the following while a more detailed description can be found in Appendix C.3.

The remote sensing data consists of PlanetScope imagery (Planet Team, 2017) from 19 dates with a pixel size of 3 m. The number of days between the first and last imagery is 192 days (24-04-2017 – 02-11-2017). The observations extracted from the remote sensing data provided a time series of the area of surface water in each depression on the 19 dates over the dry season in 2017. To extract this time series, the NDWI is calculated (McFeeters, 1996) and automatic thresholding (Otsu, 1979; Rosin, 2001) and a trajectory analysis (Powell et al., 2008; Zomlot et al., 2017) is applied. The depressions all run dry during the dry season, albeit at different times; the observation of interest is the last day of water inundation.

In a seismic refraction survey, the travel-time of the compressional P-wave from a seismic source to a series of known receiver locations is measured. The spatial distribution of the first arrival travel-times is controlled by the velocity below the profile. The travel-times can be inverted for using travel-time tomography to generate a 2D profile of seismic velocities (e.g. Zelt et al., 2013), but in this application we use the travel-times directly. For this study, we used 24 geophones spaced at 4 m resulting in 92 m profiles. The seismic source was a sledgehammer swung onto a 20 cm by 20 cm steel plate. The shot spacing was 4 m. To increase signal-to-noise ratio, 6 shots were stacked at each location. Since all the sinkholes were longer than 94 m, each seismic line was rolled at least once with the succeeding line starting in the middle of the first line at four depressions (S1 to S4) resulting in a total line length of 140 m. In the case of S5 the profile was rolled twice resulting in a total line length of 184 m. The first arrival time of the P-wave were picked manually. Noisy traces where the first arrival was not clear were not picked. The reciprocal travel-times, which represent the energy traveling from A to B and then again from B to A, were used as an indication of the uncertainty of the observations.

4.4.4 Forward modelling

As indicated in Figure 4.1, the forward model for the process structure is a bucket water balance model, while for the physical structure it is a shortest path raytracing model combined with a rock physics model. A short overview of the forward model functioning is given in the following while a more detailed description can be found in Appendix C.4. Further the prior parameter distributions are also defined in Appendix C.5.

The response from the alternative process structures is obtained from a bucket water balance model for accounting of the water presence in the depressions over the dry season. The forward model calculates the duration of water inundation after the start of the dry season. Four different models are defined based on the alternative process structures (Figure 4.3) defined in Section 4.4.1. The input to the forward models consists of samples from prior parameter distributions of maximum water depth in the individual depressions and rates of evaporation, groundwater recharge and groundwater discharge.

The forward response from the alternative physical structures is obtained from a combination of a rock physics modelling and ray tracing using a shortest path algorithm (Moser, 1991; Rücker et al., 2017). Instead of picking a velocity range for each facies in our physical structure models we elected to utilize rock physics relationships to provide reasonable estimates for the velocities expected by our structure. Following the methods of Flinchum et al. (2018) and Holbrook et al. (2014) we estimate the seismic velocity of a porous media by using a Hashin-Shtrikman relationship to define velocities as a function of porosity and then us Gassmann’s equations (Brie et al., 1995; Mavko et al., 2009) to adjust for water saturation.

Once the model structures were transformed into velocity with the refraction class from the open-source Python framework pyGIMLi (Rücker et al., 2017). The travel-times are based on shortest path methodology (Moser, 1991) of seismic waves. The velocity sections that are based on the physical structures are then passed to pyGIMLi to produce the forward modelled travel-times. These travel-times are what are used in the subsequent analysis below. The input to the forward models consists of samples from prior parameter distributions of geometric (controlling the boundary locations in the physical structure) and lithological (controlling the velocity structure) parameters.

4.4.5 Model rejection

In the model rejection step (step 4, Figure 4.1), the two independent data sets (remote sensing and seismic refraction) are used to reject the alternative models of process structure (using remote sensing) and physical structure (using seismic refraction).

For the time-dependent surface water area of a depression extracted from the remote sensing data, a likelihood function is defined (Appendix C.3). The likelihood function is defined as a beta distribution that has a bounded interval describing possible values. This means realisations from the forward model that lie outside of the bounds are assigned a zero likelihood. Model structures obtaining a zero marginal likelihood are rejected.

To globally test the physical models, we apply a dimension reduction technique to the refraction seismic data. We apply Principal Component Analysis (PCA) using the PCA class from the open-source Python framework scikit-learn (Pedregosa et al., 2011). PCA reduce the number of dimensions by feature extraction, which means that from the first arrival variables (>900) we can create a smaller number of new variables that are combinations of the old variables. These new variables are ordered by how much variability they explain of the old variables. In our case we only keep the two first principal components as they explain more than 90 % of the variability. The principle components do not explain anything physical, but they describe the variability of the input data.

Based on the first two components of the PCA, we apply a novelty detection analysis (Markou and Singh, 2003) using the LocalOutlierFactor class from scikit-learn (Pedregosa et al., 2011). Novelty detection evaluates whether a new observation in a dataset is an outlier or not. In this case we evaluated whether the observation dataset (in terms of its first two PCs) is an outlier relative to the synthetic forward model response. If the observation dataset is deemed to be a novelty (i.e., outlier) relative to the model response, the model can be rejected. Similar approaches to this has been applied in other studies (Hermans et al., 2015; Park et al., 2013; Peeters et al., 2013; Pirot et al., 2019; Scheidt et al., 2015).

4.4.6 Update model probability

4.4.6.1 Process structure models (Y_{ps})

The model evidence for the individual process structures is quantified by the marginal likelihood of the observed data given the model structure and parameters. The likelihood function for the depressions being dry is defined as a beta distribution based on the values from the time series of surface water area in the depressions (Section 4.4.3). A more detailed description of the derivation of the likelihood function can be found in Appendix C.3.

4.4.6.2 Physical structure models (Y_{pl})

To assign probabilities to the proposed physical structure models, we use a logistic regression classifier using the LogisticRegression class from scikit-learn (Pedregosa et al., 2011). Logistic regression is a supervised machine learning technique that can probabilistically classify data into discrete outcomes.

The classifier is trained on the response from the forward geophysical model (i.e. the model realisations of first arrival data) using a known physical structure and is then applied to the observation data to predict the probability of each physical structure.

4.5 APPLICATION RESULTS

In this section, we present the results of the applied methodology to the conceptualization of the depressions in Wildman River area. Our goal is to assess the plausibility of different conceptualizations by testing both the process structure and the physical structure with two different, and more importantly, independent types of data. The response of the forward models, a geophysical model and a water balance model, is compared with the observation through the model rejection step and probability update step (Figure 4.1, and Section 4.5.1 and 4.5.2). Last, the updated probabilities of the process structure and physical structure models are combined to assess whether the depressions are connected to groundwater (Section 4.3).

4.5.1 Process structure

The water balance forward model (step 3 in Figure 4.1) generated 4,000 synthetic data sets of surface water area response of each of the five depressions (S1 to S5) from which the days the depressions are no longer inundated have been obtained. For each process structure (I to IV), 1,000 responses are available, which represent prior beliefs given the prior parameter values and process structures. The cumulative probability of days, since the end of the wet season, at which the depressions are no longer inundated, is shown in Figure 4.3 (shaded areas). The x-axis is limited to 250 days, as this is the maximum length of the dry season. The cumulative likelihood distributions (defined in Section 4.4.6.1) for the individual depressions, based on the remote sensing data of open water surface area, are indicated with a solid black line.

In the model setup we have assumed that groundwater recharge and discharge effectively cancel each other out in a water balance, and therefore the results from model structure I and IV are the same. This non-uniqueness problem arises as the data will not be able to distinguish whether the water in the depression is a local (structure I) or regional feature (structure IV).

The presence of a groundwater recharge component in structure II decreases the number of days to dry out the depressions, while the presence of a groundwater discharge component in structure III increases the number of days required to dry out the depression.

The results from the four different structures (I to IV) in depressions S1, S2 and S5 look similar, with the first realisations drying out after around 100 days. The realisations pertaining to S3 and S4 dry out at a slower rate, with none of the process structures in S4 drying out before the end of the dry season. This can be attributed to the greater maximum depth (Appendix C.5) of S3 and S4 and therefore higher starting volumes.

Comparing the model responses (coloured histograms) to the likelihood function based on the remote sensing data (solid black line in Figure 4.4) reveals that the depressions are drying out much faster in reality than most of the forward model responses. Of the simulated responses, process structure II (green area) displays the fastest drying out curve for all five depressions, while process structure III consistently maintains water in its depressions for a considerably longer period.

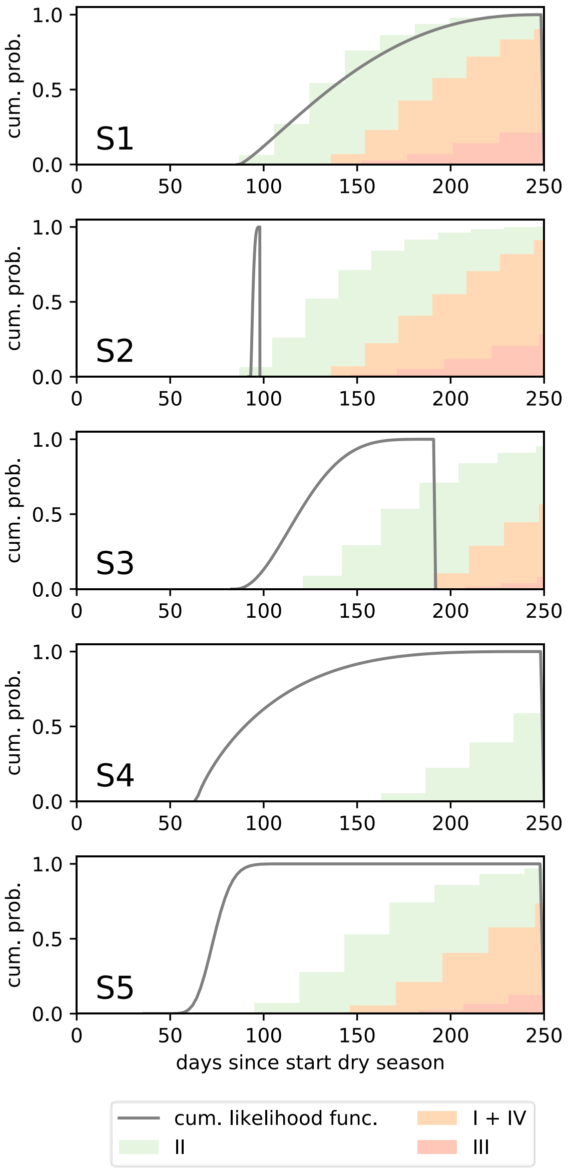

Figure 4.4. Cumulative probability (cum. prob.) of the day after the start of the dry season where the depressions are no longer inundated from realisations of water balance models given different process structures I-IV (Figure 4.3). The cumulative likelihood function (cum. likelihood func.) shown in black, is defined based on remote sensing data. If all realisations of a process structure plot outside the lower and upper limit of the likelihood function, the process structure is rejected.

The performance of the different process structures (I to IV) for the different depressions (S1 to S5), given the remote sensing data, is further discussed in Table 4.2. Here we show the probability values based on marginal likelihoods of the individual models obtained from the likelihood function shown in Figure 4.4. Structure II clearly shows the smallest discrepancy with the observations and is outperforming the remaining process structures in all depressions. Any zero probability values indicate process structures whose realisations are all outside of the data-derived likelihood function and therefore obtain a zero marginal likelihood. Such model structures are rejected (step 4 in Figure 4.1), while the probability of the remaining model structures are updated (step 5 in Figure 4.1). In depressions S2 and S4, only process structure II remains plausible after the testing exercise, while all four process structures are still plausible for S1 and S5.

Table 4.2. Probability of different process structures I-IV (Figure 4.3) given the remote sensing data. The reported probabilities are marginal likelihoods of 1,000 synthetic forward model runs based on a likelihood function defined based on remote sensing data. Grey boxes indicate that the observation (likelihood function) is outside the prior range obtained from the forward water balance model.

	I + IV: no connection + flow-through	II: Groundwater recharge	III: Groundwater discharge
S1	0.50	0.37	0.13
S2	0.00	1.00	0.00
S3	0.01	0.99	0.00
S4	0.00	1.00	0.00
S5	0.37	0.56	0.07

4.5.2 Physical structure

The geophysical forward model (step 3 in Figure 4.1) generated 4,000 synthetic data sets of seismic responses for each of the five depressions (S1 to S5). For each physical structure (A to D), 1,000 responses were available. The interpretation of these datasets by means of principal component analyses is discussed on the basis the first two principal components (dimensions) shown in Figure 4.5. The first two dimensions represent between 87 % and 93 % of the total variance of the seismic response for each of five depressions. In each plot, the red cross represents the observed response. The five rows of Figure 4.5 present the result for the five depressions (S1-S5), while the four columns present the effect on the seismic response of applying different physical structures (A to D).

The input to the rock physics model for generating the seismic responses is the same for all depressions, therefore the difference between depressions S1 to S5 is solely due to the different inputs to the structural model. For depressions S4 and S5 more variation between forward model response is evident than for the three other depressions (note the different axes on Figure 4.5). This is probably due to the uncertainty of the water table that is relatively well known in the S1 to S3 depressions but less well known for S4 and S5 (supplementary information).

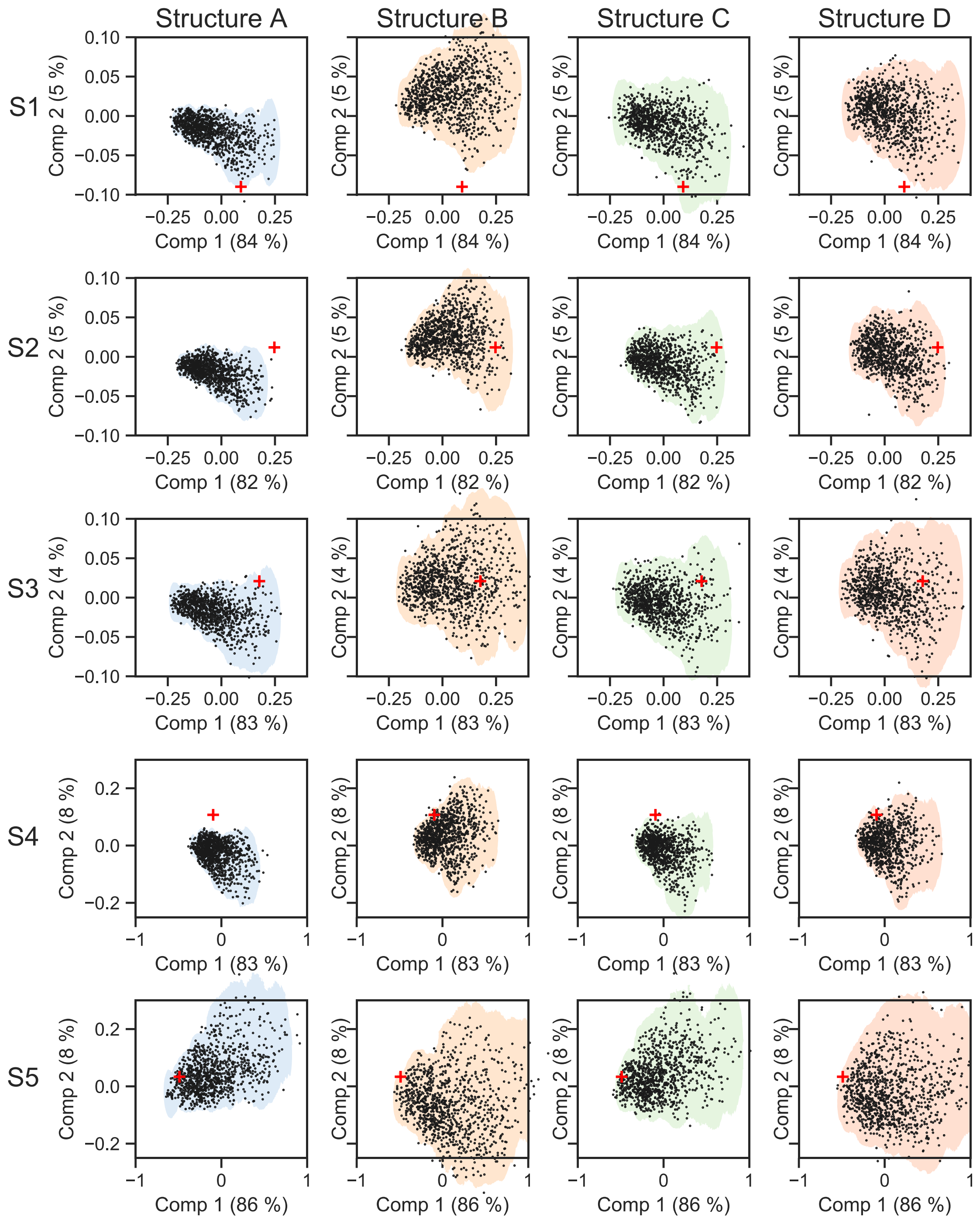

Figure 4.5. First and second component of the PCA of the forward model response of the geophysical models for depression S1 to S5 (rows) for each physical model structure (A to D). Note the different axes for S1 to S3 and S4 to S5. The coloured area indicates the frontier outside which the addition of a datapoint would be classified as a novelty. If the observation (red cross) plots outside the shaded area, a novelty detection algorithm classifies the observation as a novelty in relation to the model realisations and the model structure can be falsified.

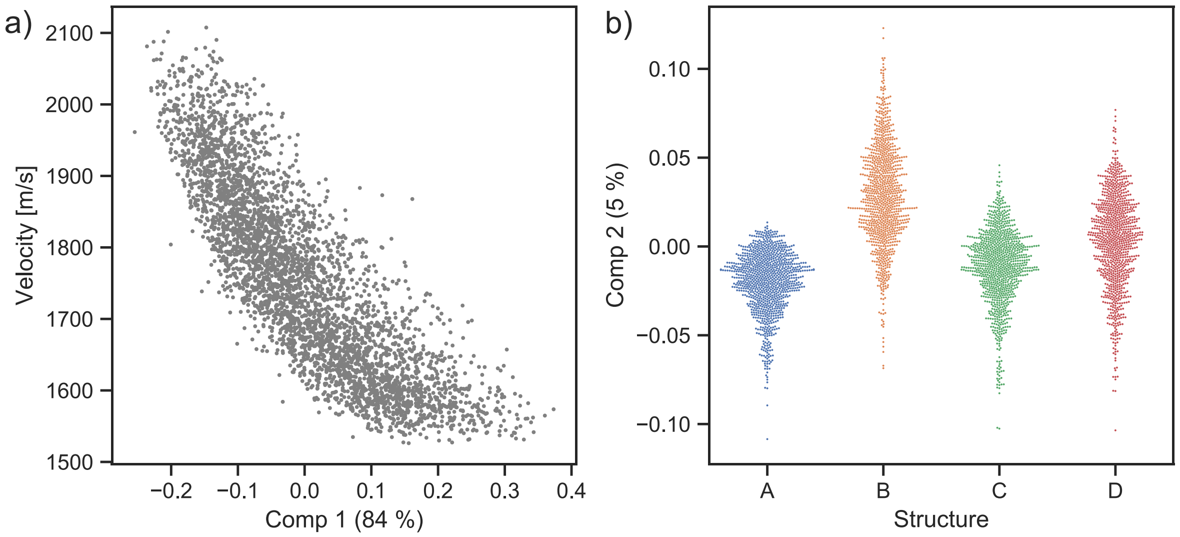

Figure 4.6. a) Relation between the first PCA component and velocities for S1 and b) the second PCA component and the applied physical structures for S1. The illustrated velocities are mean velocities in the saturated sand zone. The first PCA component generally explains the differences in velocity, while the second component explains the differences in model structure. This means that the alternative structures do influence the forward response, but the effect is secondary to the velocity influence.

The relationship between velocity (derived from the simulated seismic response) and the first PCA component and between different model structures and the second PCA component have been plotted for S1 (Figure 4.6). The velocity is seen to decrease with increasing values of the first PCA component. The second PCA component has lowest values for physical structure A, intermediate values for C and D and highest values for structure B. This pattern also holds true for the rest of the depressions (except for S5 where the relation between structure and the second PCA component is reversed). Figure 4.6 illustrates that the first component generally explains the velocity, while the second component explains the effect of different physical structures.

Physical structure A and C are shown to generate a similar response with relatively lower variation than B and D; note that the latter two structures also show a similar response. A model which has a higher porosity zone (B and D) therefore has larger influence on the response (more positively correlated with the second principal component), while the addition of a clay layer in the structural model (A and C) does not change the result significantly (low and negative correlation with second principal component).

The observed data is illustrated with a red cross on Figure 4.5. Observation uncertainty has been considered (but not plotted) and generally leads to a variation of less than +/- 0.005 for the first two principal components. When the observed data point plots outside the forward model response (the shaded area), the data point cannot be predicted by the prior. A novelty detection analysis is applied to automatically classify the observation data as a novelty or an inlier within the prior range (step 4 in Figure 4.1). Recall that when the observation data is classified as a novelty, the prior used to generate the synthetic data is rejected. The coloured cloud indicates the frontier between inliers and outliers derived from the model realisations. The rejected models are assigned a zero probability in Table 4.3. For the accepted models, probabilities are calculated with logistic regression classification (step 5 in Figure 4.1).

The results of logistic regression classification of the observed data trained on the realisations is shown in Table 4.3. Some correspondence can be observed between the observed data location on Figure 4.5 (red cross) and the probability performance of the different physical structures, i.e. when the response from a model structure is densest around the observation, it performs better. Note, however, that while only two dimensions of PCA are shown in Figure 4.5, the logistic regression is based directly on the simulated values (thus uses the entire data set).

For depression S1 and S2 the posterior probabilities are similar, indicating the alternative structures (A and C for S1 and B, C and D for S2) do not generate responses different enough to be able to discriminate between them. Depressions S3 and S4 shows a slight preference for structure B (0.37 and 0.67 probability, respectively), while S5 shows a preference for structure C (0.52 probability) and generally low preference for structure B and D, including the high porosity zone in the middle of the depression.

Table 4.3 . Model probabilities assigned through logistic regression classification of the observed data. Models rejected through novelty detection (grey) are assigned a zero probability.

	A: homogeneous	B: high porosity zone	C: clay layer	D: high porosity zone + clay layer
S1	0.48	0.00	0.52	0.00
S2	0.00	0.33	0.33	0.34
S3	0.10	0.37	0.25	0.28
S4	0.00	0.67	0.00	0.33
S5	0.29	0.07	0.52	0.12

4.5.3 Posterior probabilities of conceptual models

Comparing the posterior probabilities obtained for the process structure (Table 4.2) and the physical structure (Table 4.3), it is apparent that the process structure probabilities are more decisive. This demonstrates the importance of the sensitivity of the datatype towards the developed alternative model structures and its ability to discriminate between them. Figure 4.6 illustrates that the physical structure affects the realisations (based on PCA component 2), but that it is secondary (i.e. lower correlation) to the velocity estimates (based on PCA component 1). The velocities are estimated based on many parameters with relatively wide priors as they are based on literature values (Appendix C.5). In order to resolve and better discriminate between structures, informative priors are needed. On the other hand, while the difference between the process structures are more pronounced, despite the larger uncertainty of the observations, we can better discriminate between the models.

The posterior probability of the conceptual models of the five depressions combined from the probability of the physical structure and the process structure are shown in Figure 4.7. The posterior probability is relatively decisive, especially in S2 and S4. It stands out that the different depressions obtain quite different results, although the expectation was that the depression would behave similarly. This illustrates the importance of considering more than one conceptual model and to have a sufficiently large testing data set.

Of the 16 different conceptual models, only two would allow for the depression to act as conduits of recharge, A-II and B-II. These models have been marked with a dashed box on Figure 4.7. In depression S2, S3 and S4, conceptual model B-II is outperforming the rest of the models indicating consistency with the depressions being conduits of recharge. In S1 the observations plot on the edge of the prior range of realisations for all physical structures (Figure 4.5), though structure A and C are still classified as probable models. This indicates that the "real" physical structure might still be an unknown unknown. Depression S5 is least decisive in terms of the model structure for both the physical structure and the process structure, and more model testing is required to discriminate between model structures.

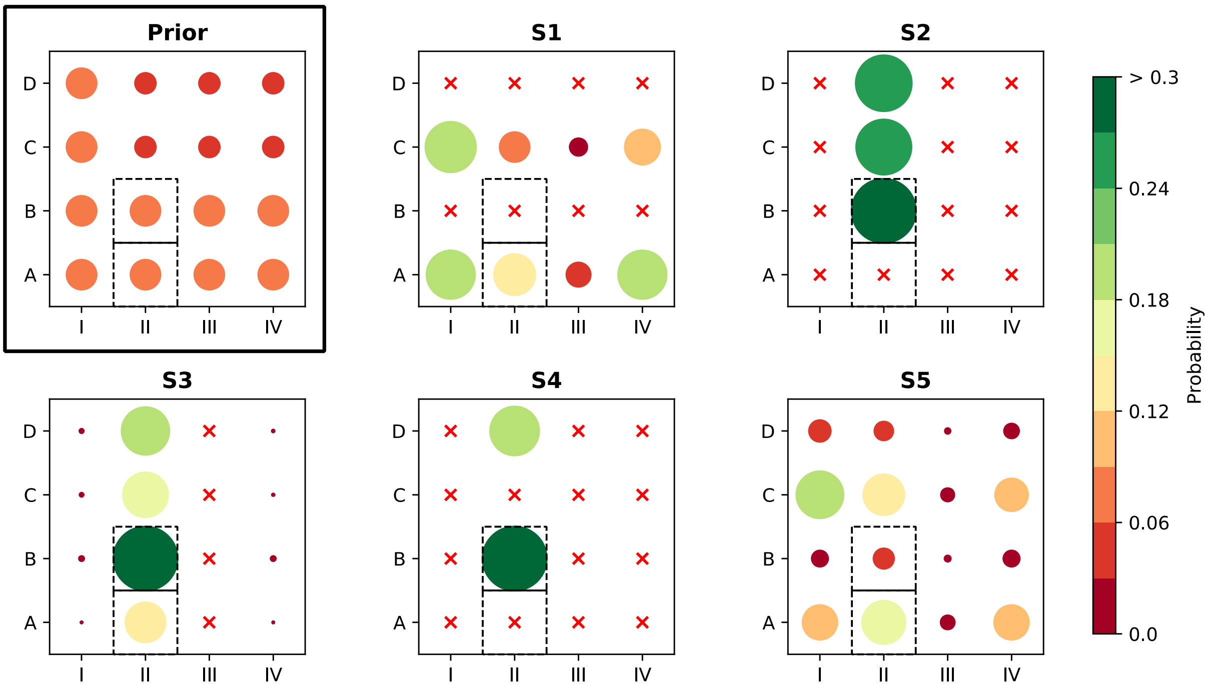

Figure 4.7. Posterior probability of the combined physical (A to D) and process structure (I to IV) for the individual depressions (S1 to S5). The prior probability for all depressions is shown in the top left plot. Dashed boxes identify the models that would allow for the depression being conduits for recharge. The conceptual models that have been marked with red crosses are rejected.

4.6 DISCUSSION

The combined updated model probabilities of the physical structure and the process structure revealed, based on independent data, that the depressions act as conduits for recharge for three (S2 to S4) out of five depressions. For the two other depressions (S1 and S5), the data is more indecisive, and more testing would be needed to discriminate between model structures.

We applied a systematic testing workflow consisting of:

- Using a Bayesian framework with a rejection step that compared performance to data.
- Using data for testing rather than development.
- Using two lines of independent evidence for conceptual model testing.

By applying this workflow rather than an inversion approach, we were able to uncover more than just one conceptual model that is consistent with the data for all depressions. It is also worth noting that the model testing exercise has changed our understanding of the depressions considerably. This is illustrated by comparing the prior probability to the posterior probability.

Multiple lines of evidence are almost always used when developing conceptual models e.g. (Banks et al., 2019; Bresciani et al., 2018), hence the combination of geophysics and remote sensing data to develop conceptual models is not novel e.g. (Francés et al., 2014; Othman et al., 2018; Youssef et al., 2012). However, using multiple lines of evidence in a conceptual model development approach where data is used for testing rather than development is still rare. By integrating multiple lines of evidence in this study we gain more confidence in the conceptualization of the depressions. Also, in a Bayesian workflow, more model testing dilutes the effect of the choice of prior model structure probabilities, whose definition is the most controversial component in the Bayes framework (Sambridge et al., 2013).

In line with findings of other model ensemble studies (Højberg and Refsgaard, 2005; Rojas et al., 2010c; Seifert et al., 2012), several conceptual models are consistent with current knowledge and observations. The ensemble of different model structures is obtained by using data to test models rather than developing them as in the consensus approach. Disregarding alternative plausible model structures can eventually lead to biased and overconfident predictions.

As for the model testing exercises in e.g. (Hermans et al., 2015; Scheidt et al., 2015; Schöniger et al., 2015a), we have applied our testing approach in a Bayesian context, where model structure prior probabilities are updated depending on model performance. The Bayesian approach provides a formal framework for iteratively incorporating new data and insights (Figure 4.1). In each model testing step, models that are inconsistent with data can be removed from the ensemble and conceptual surprises are thereby accommodated.

The insight into the system functioning gained from testing alternative conceptual models can be used in future modelling exercises. With more confidence in the conceptual model, we have more confidence in predictions of (future) system behaviour, which provides more robust evidence to underpin environmental management decisions.

Any multi-model approach is limited by our inability to define a collectively exhaustive range of models. However, by applying a rejection step in which model structure performance is compared to data rather than having an intercomparison of performances of different model structures, we avoid assuming the true model is within the ensemble of models tested. Indeed, following the workflow from Figure 4.1, all models can still be rejected. However, as probability adds up to unity after updating the probability, we are implicitly assuming the range of models is collectively exhaustive.

Another limitation of the testing approach is that the result is limited by the information content in the testing data. When the information content is low, the definition of the prior probability becomes very important for the result of the testing exercise (Rojas et al., 2009). In relation to the information content of the data, the remaining plausible models after a model rejection step are only *conditionally accepted* because they have not been proven to be inconsistent with data yet (Oreskes et al., 1994).

The results are also limited by our assumption that the success and failures of model realisations relate to the model structures. Other uncertainties that may give rise to false positives (not rejection inconsistent model structure) or negatives (rejecting consistent model structure) include prior parameter definition, mathematical translation of the model structures and forward model definition. The testing exercise can only tell us if some part of hypothesis/forward model data is not right but not which part.

Nevertheless, our approach is generic and can be applied to any hydrogeological conceptual model where conceptually uncertain component(s) exist. However, the generalizability of the methodology is limited by the complexity of the forward model. While running forward models is less intensive than model inversion, the number of model realizations needed to obtain a reliable model probability in Step 5 of our workflow increases the computational burden.

Finally, the applied approach also allows us to uncover conceptual surprises (in case one would reject all models in the model rejection step), but it does not tell us how to deal with them. A conceptual surprise prompts the development of new hypotheses based on model behaviour. These would, at least indirectly, be based on the former model testing data, and therefore it should not be used for model testing again to avoid circular reasoning. It is beyond the scope of this paper to address this issue.

4.7 CONCLUSION

We proposed a systematic approach to hydrogeological conceptual model testing, which is needed to increase transparency in the groundwater modelling workflow and seek out conceptual surprises.

The approach focuses on using independent data to test models, rather than to develop them. Also, we have emphasised that models should be rejected by comparing performance against data rather than comparing model performance against competing models.

We have applied the approach to the Wildman River Area, Australia, involving the testing of the connectivity of widespread, enclosed depressions to groundwater.

Our suggested approach is generic and can be applied to any hydrogeological conceptual model where conceptually uncertain component(s) exists.

Chapter 5: Thesis Conclusion and Outlook

In order to manage groundwater resources effectively, we need a good understanding of the groundwater system in question. This thesis has focussed on methods for improving the understanding of the conceptual model that formalises and integrates many of the underlying assumptions in a groundwater model. The conclusions that can be derived from this thesis can be divided into contributions to the general methodology for hydrogeological conceptual model development and testing and contributions to the hydrogeological conceptualization of Wildman River area. Finally, an outlook will be provided, discussing some future research directions.

5.1 GENERAL METHODOLOGY

5.1.1 Model development (Research aim 1)

In this thesis we have advocated for a systematic approach to conceptual model building, but the act of discovery is not always seen as a scientific logical process (Schickore, 2018). The discovery of a model in the consensus approach in hydrogeology is somewhat guided by rules, e.g. (Barnett et al., 2012; Brassington and Younger, 2010), however this is not the case for the multi-model approach. This has led to a wide variety of approaches in literature around multi-modelling (Chapter 2). We identified current approaches and unified scattered insights to apply best approaches. We found that a hypothesis testing approach consisting of mutually exclusive alternative models was most suitable to develop alternative models.

The conceptual model consists of physics-based hypotheses concerning the process structure and the physical structure. The process structure includes dominating processes and it is often assumed to be relatively well-known (Carrera et al. 1993). For example, the alternative understandings of groundwater-surface water interactions between an aquifer and a lake will be limited to the lake being: 1) a groundwater recharge feature, 2) a groundwater discharge feature, 3) a flow-through feature, or 4) not connected to the groundwater. These different understandings are based on a literature review of the general understanding of surface water-groundwater interactions. When characterising the process structure, the difficulty of the problem arises to a higher degree when characterising the spatial variability of the controlling parameters. This is however, not related to the conceptual model.

For the physical structure the alternative conceptual models are generally considered to be more difficult to characterise. The approaches have therefore in literature been more varied. Alternative conceptual understanding of the physical structure could test the existence of palaeovalleys, faults, confining layers, continuity of layers etc. The exact location of these features would be part of the lower order uncertainty, i.e. not the conceptual model.

As practical applications we developed models based on literature reviews. In Chapter 3 we developed alternative models starting from an existing conceptualisation and identified key assumptions in an initial water balance that needed systematic testing. In Chapter 4 we developed alternative models of sinkhole-like depressions and their role in groundwater-surface water interactions based on general principles and understandings of such systems. Further, we emphasized the need to set up alternative models in a factorial approach to ensure that assignment of performance can be linked to the appropriate hypothesis that is being tested. Also, in both chapters bold hypotheses were considered, challenging what we considered to be plausible initially. By developing bold models, we were minimizing the risk of conceptual surprises.

5.1.2 Model testing (Research aim 2)

A literature review in Chapter 2, showed that model selection techniques are often applied rather than model testing techniques in the multi-model approach. This may be due to the requirement that testing data must be independent, while model selection data does not. We advocated for a systematic testing approach in Chapter 2 and 4 to increase confidence in conceptual models.

In our application studies, by limiting the use of data in the model development process, we attempted to avoid confirmation bias in the initial model ensemble and thereby gained a relatively wide prior range of conceptual models. Further, in Chapter 4, this approach led us to put aside as much data as possible for model testing. Being able to use data for testing conceptual models rather than development increases the confidence in the models as it gives a possibility to justify the ensemble. Further, more data to test models helps to better discriminate between the alternatives and thereby dilutes the effect of the prior probability. Lastly, using more data for model testing increases the chance of model rejection, which in the end is even more valuable than justifying a model as it explains why that specific model can be excluded from the model ensemble. In Chapter 4 we showed that it is not only the uncertainty of data or the number of datapoints that is important for the discriminatory power of the data, but also how sensitive the observation is to the alternative model structures.

In Chapter 3 we only used the closure of the water balance to test models and were not able to reject any of the conceptual model structure combinations. We therefore further applied a Bayes Factor to evaluate whether some models were preferred over others. In Chapter 4 we had reserved more data for testing, including remote sensing and refraction seismic data, and were able to reject a few models. The testing approach provided us with a transparent and reproducible explanation of why the rejected models were not considered plausible.

As the characterisation of conceptual uncertainty relies on the definition of multiple model structures rather than just a single model structure, it leads to increased cost and CPU time required. Therefore, this thesis has focussed on model testing using simple models rather than complex 3D numerical models to make model testing more accessible. The alternative to using simple models is the application of surrogate modelling if more complex models are deemed necessary. However, surrogate models are not applied in this thesis. Chapter 3 illustrated the use of simple stochastic water balance models to test competing conceptual models. Chapter 4 applied the testing procedure in a hydrogeological characterisation using a forward geophysical model and a bucket water balance model. By applying simple forward models, we have shown that model-based conceptual model testing does not have to be a time-consuming task.

5.1.3 Impact on predictions

Multiple models are difficult to develop, test and run but also more difficult to explain and present. Presenting the predictions from multiple models rather than from a single deterministic model inevitably becomes more complex. Most groundwater modelling studies employ a deterministic approach so understanding and communicating papers that employ the multi-model approach might appear both ambiguous and confusing. The results presented in the applications studies in this thesis consisted of presenting ranges and have acknowledged that unknown unknown conceptual models may exist. This type of communication of results might appear a bit more confusing than just presenting the results of a deterministic model, but it is a more honest representation of the uncertainties involved in groundwater modelling.

5.2 WILDMAN RIVER AREA (RESEARCH AIM 3)

The methods developed in this thesis were applied to the study site in Wildman River area. In Chapter 3 a stochastic water balance framework that incorporates the outstanding conceptualisation questions for the Wildman River area was developed. This study confirmed and provided more confidence in the results of an existing (Tickell and Zaar, 2017) deterministic water balance. In chapter 4 we sought to increase the understanding of sinkhole-like depressions that were hypothesized to act as conduits for recharge (Tickell and Zaar, 2017; Turnadge et al., 2018a). Three out of five depressions that were used as a test case were conditionally confirmed to act as conduits for groundwater recharge, while for the last two depressions, the data was inconclusive. The study highlighted that not all sinkhole-like depressions act as sources of recharge and that it is not possible to assess the connectivity to groundwater solely from remotely sensed observations.

5.3 OUTLOOK

In chapter 4 we applied an approach where almost all data was put aside for model testing where the developed models were developed based on a literature review of general knowledge about sinkholes and groundwater-surface water interactions. However, when the problem becomes more complex, the task of developing alternative models based only on a general literature review, may not be manageable. The number of plausible models will be too many. The problem with using data for model development is that the range of developed models is subject to confirmation bias. A method where data can be used systematically to develop alternative models while avoiding confirmation bias is still needed.

In chapter 3 the only data used for model testing was the closure of the water balance. This limited the possible discrimination between alternative models. In Chapter 4 we used both refraction seismic data and satellite imagery and were able to reject several conceptual model structures. The data used in testing exercises reported in the literature has mainly been with geophysical data. Avenues of model testing with other hydrogeological data such as pumping test data and environmental tracers in a Popper-Bayes approach has yet to be explored.

Although parts of this thesis have been devoted to discussing conceptual surprises and how to uncover them, we have not discussed how to deal with them. A conceptual surprise should prompt the development of new hypotheses based on model behaviour, but in literature a reaction to model rejection has sometimes been an ad-hoc modification, assuming something is wrong with the mathematical representation or due to simplifications, not the underlying conceptual model. If a new model or range of models are developed based on model rejection, it would at least indirectly be based on the data used for model testing. Therefore, the model testing data should not be used for model testing again to avoid circular reasoning. A framework of how to deal with conceptual surprises is still needed.

Appendix A: Chapter 2

A.1 MODEL DEVELOPMENT LITERATURE REVIEW

Table A.1. Examples of approaches to develop conceptually different models for the Conceptual Physical Structure (Ph), Conceptual Process Structure (Pr) and the Spatial Variability Structure (SVS). Approaches to developing different models include hypothesis testing (H), complexity testing (C) and interpretation testing (I), i.e. Figure 2.3. If the model objective is defined in the introduction of the paper the objective of the model is here considered well defined. The model objective is relevant to this table as the model objective should have an impact on what to include in the conceptualization.

Study	Is the model objective well defined?	Conceptual multi-model development approach	Ph	Pr	SVS
Altman et al. (1996)	Yes	Two different representations describing unsaturated zone flow through fractured media including equivalent continuum and a dual permeability model.		H	
Aphale and Tonjes (2017)	No	Top of semi-confining unit either as uniform surface or undulating based on interpolation between boreholes (H). Northern extent of semi-confining unit represented by two different models (H). Vertical discretization of downward fining sediment in aquifer as either uniform or variable (H). Landfill effect on recharge either (i) no effect on recharge, (ii) recharge diverted to recharge basins adjacent to the landfill mounds, (iii) all recharge collected for off-site treatment (H). Drains segmented or not (H).	H	H	H
Carrera and Neuman (1986)	No	Ten alternative zonation patterns of hydraulic conductivity for synthetic aquifer.		C	
Castro and Goblet (2003)	Yes	Four alternative models where constraints within a formation is imposed (i.e., linear, exponential or with increasing distance decrease in hydraulic conductivity or constant hydraulic conductivity values for all formations).		H	
Elshall and Tsai (2014)	No	Two different geological formation dips propositions (H). Three indicator geostatistical methods for representing geometry: indictor zonation, generalized parameterization and indicator kriging (H).	H	H	
Engelhardt et al. (2014)	No	Seven alternative conceptual models varying the number of parameters (horizontal and vertical hydraulic conductivity and specific yield) in 10 homogeneous zones by lumping zones together.		C	
Feyen and Caers (2006)	Yes	Two different training images representing two different braiding and sinuosity scenarios of a fluvial system (H). Three different affinity and angle maps representing local variation in channel width and orientation (H). Three different variogram types: spherical, exponential or Gaussian (H).		H	
Foglia et al. (2007)	No	Five alternative models that differs in zonation of hydraulic conductivity. Alternatives developed by lumping together different zones of homogeneous hydraulic conductivity.		C	
Foglia et al. (2013)	Yes	Two different bedrock geometries defining the bottom of the groundwater system based on different data (I) Five different zonation of hydraulic conductivity (C). Recharge either zero, spatially uniform, zonated based on soil types or simulated through rainfall-runoff model (I).	I	C	I/H

Study	Is the model objective well defined?	Conceptual multi-model development approach	Ph	Pr	SVS
		Streams are described with MODFLOW's SFR and River package in alternative models imposing different assumptions (H).			
Gedeon et al. (2013)	Yes	An initial model including a crude description of e.g. a clay aquitard and an update of the initial model including new information to update the description of the aquitard. This is an example of a consensus approach allowing for updates and the classification system presented by Figure 2.3 therefore does not apply.	N/A	N/A	N/A
Harrar et al. (2003)	Yes	Two manually created alternative geological models are based on the same data and contains the same five sediment types but is interpreted by two different geologist. They differ in regards to the way the sediment type is assigned to the cells based on borehole data and the number of layers. Thereby one model reflects a more heterogeneous system while the other reflects a stratified system.	I		
He et al. (2014)	No	Two training images for an MPS algorithm where one is based on SkyTEM data and the other is based on a Boolean simulation.		H	
Hermans et al. (2015)	Yes	In the field example four different training images are produced through a Boolean simulation for an MPS algorithm to describe variation between sand, clay and gravel.		H	
Hills and Wierenga (1994)	Yes	Unsaturated zone and transport models developed by five different teams. The models differed in regards to soil being modelled as isotropic or anisotropic and homogeneous or heterogeneous.		I	
Højberg and Refsgaard (2005)	Yes	Three hydrogeological models manually generated by three different teams for different purposes.	I		
Johnson et al. (2002)	Yes	A one-layer, two layer and three layer model is considered to represent a layered basalt and interbedded sediment aquifer.	H		
Kikuchi et al. (2015)	Yes	Inclusion of zero, one or two lenses of higher hydraulic conductivity in an otherwise homogeneous unconfined aquifer (H). Mountain front recharge as either a continuous line parallel to mountain front or through discrete stream features (H). Two models with and without underflow through subsurface zone to adjacent basin (H).	H		H
Knopman and Voss (1988), Knopman and Voss (1989)	Yes	Input of solute at upstream boundary of either i) constant, ii) decaying or iii) spatially varying initial condition (H). Two different models in regards to whether first-order decay is affecting the transport (H). One or three layers to describe the medium of well-sorted sand and gravel (C)		C	H
Knopman et al. (1991)	Yes	One-dimensional models of solute transport differing in regards to whether first-order decay is affecting the transport (H). One, two or three layer to describe the medium of well-sorted sand and gravel (C)		C	H
La Vigna et al. (2014)	Yes	Three models considered to explain connection between two sand aquifers is i) outside of groundwater model, ii) through silty-sandy lense and 3) through old, not backfilled well.	H		
Lee et al. (1992)	Yes	Homogeneous, layered and randomly heterogeneous geologic description to model tracer migration.		C	
Li and Tsai (2009)	Yes	In the Baton Rouge Area case study: Three different influences of a fault in regards to connectivity between aquifers is considered: i) impermeable fault model, ii) low permeability model and iii) no fault model.	H		
Linde et al. (2015)	No	Two training images for an MPS algorithm where one is based on a local outcrop and the other is based on an aquifer analogue.		H	
Lukjan et al. (2016)	No	Two hydrogeological interpretations, homogeneous or zoned (C). Five models by combining different outer boundary conditions as either head or no-flow boundaries (H).		C	H

Study	Is the model objective well defined?	Conceptual multi-model development approach	Ph	Pr	SVS
Mechal et al. (2016)	No	Two different models with two different fault sets and one model not representing faults at all (H). Five models with increasing number of transmissivity zones (C). Two models with one representing all rivers and one only representing the major river (C). Two models of lateral boundary conditions where one considers outflow to an adjacent aquifer and one does not (H).	H	C	C/H
Meyer et al. (2003)	No	Nine different variogram models to explain log air permeability variation in unsaturated fractured tuff.		H	
Meyer et al. (2007)	Yes	Two alternative models of spatial distribution of K: Homogeneous and zoned. A steady-state and a transient boundary condition to a stream.		C	C
Nettasana (2012)/Nettasana et al. (2012)	No/Yes	Three/two different independent interpretations of geology that differ in regards to e.g. number of layers (I). Two different zonation of recharge based on either soil type, or soil type and land use (C). Two models where some lateral boundaries are either no-flow or head boundaries to test outflow to adjacent aquifers (H).	I		C/H
Nishikawa (1997)	Yes	Two models of different geometry where in the first the aquifers are horizontally layered and in the second the layers are folded offshore which would create a shorter pathway for seawater to intrude through an outcrop.	H		
Nordqvist and Voss (1996)	Yes	Three models differing in zonation of transmissivity values, i) including description of esker core and outwash material, ii) a homogeneous model, iii) including an esker core with a discontinuity and outwash material.		C	
Passadore et al. (2011)	Yes	Alternative descriptions of how aquitards pinches out in sedimentary basin affecting the connectivity of aquifers.	H		
Pham and Tsai (2015; 2016)	No	Geological description by either indicator kriging, indicator zonation or general parameterization (H). Two different fault permeability architectures: i) the same for all lithologies or ii) different for the three different lithologies (C).	H	C	
Poeter and Anderson (2005)	No	61 alternatives models by varying number and distribution of hydraulic conductivity zones generated by Sequential indicator simulations.		C	
Refsgaard et al. (2006)	Yes	In an example five different consultants are asked to assess the vulnerability of aquifers towards pollution. They solve this task with different models in terms of geometry, processes and casual relationships and end up with vastly different predictions.	I	I	I
Rogiers et al. (2014)	Yes	A geostatistical representation of an aquifer is tested against a homogeneous representation. Within the geostatistical representation 50 realization are generated representing the lower order uncertainty.		C	
Rojas et al. (2008)	No	Seven alternative representations of geometry in a synthetic study differing in regards to number of layers and which layers are spatial correlated.	I		
Rojas et al. (2010a)	Yes	Models either consider a one or a two layer hydrostratigraphic system. The hydraulic conductivity field is either described by i) constant hydraulic conductivity for each layer, ii) spatial zonation approach within the layer or iii) using Random Space Functions either conditional or unconditional. Recharge inflows originating from an eastern sub-basin described as i) diffuse recharge rates distributed over small areas of an alluvial fan, ii) point recharge fluxes at the apex of an alluvial fan or iii) recharge fluxes distributed over long sections of the eastern boundary. An additional recharge mechanism spatially distributed over the entire model domain that assumes a connection to adjacent aquifer is tested.	H	H	H
Rojas et al. (2010c)	Yes	Three alternative descriptions of geometry differing the number of hydrostratigraphic units included to test the worth of “soft” geological knowledge.	H		
Samani et al. (2017)	No	Three models consisting of different number of zones of hydraulic conductivity (C).		C	C/H

Study	Is the model objective well defined?	Conceptual multi-model development approach	Ph	Pr	SVS
		Recharge divided in four or five zones (C). Highland recharge represented by either i) a head boundary or ii) a flux boundary (H). River represented by either i) recharge boundary or ii) flux boundary (H).			
Samper and Neuman (1989)	No	Five different semi variogram models (exponential, quadratic, spherical, pure nugget and exponential with nugget).		H	
Schöniger et al. (2015)	Yes	Four alternative representations of a sandbox in a synthetic study going from simple to complex (homogenous through zonation/layered to geostatistical based on pilot points and to fully geostatistical).		C	
Seifert et al. (2008)	Yes	Two alternative model developed with and without the representation of a palaeovalley. For the study area the presence of the palaeovalley is known, but it is investigated what the impact on predicted vulnerability would be if the existence of the palaeovalley was not known.	H		
Seifert et al. (2012)	No	Five alternative hydrostratigraphic models were generated by five different (hydro) geologists in a manual approach to geological model building.	I		
Selroos et al. (2002)	Yes	Three different models describing the flow through fractured rock: i) Stochastic continuum, ii) discrete fractures, or iii) channel network.		I	
Troldborg et al. (2007)	No	Four alternative models developed different in regards to a global hypothesis about depositional history, zonation of an aquifer and which well logs to use for the interpretation.	H/ I		
Troldborg et al. (2010)	Yes	Two models that differ in regards to contact between two sand aquifers potentially separated by a clay layer (H). Two models with a different description of source zone for contamination (H).	H		H
Tsai (2010)	Yes	Experimental, spherical and Gaussian semivariogram models to describe hydraulic conductivity distribution.		H	
Tsai and Elshall (2013)	No	Three alternative variogram to explain spatial variability of the hydrofacies (exponential, pentaspherical and Gaussian) (H). One variogram applied globally or local variograms by dividing model domain in zones (C) Two fault model or one fault model dividing the model domain into three or two zones respectively (H).	H	H/ C	
Tsai and Li (2008)	No	Voronoi tessellation, natural neighbour interpolation, inverse, square distance interpolation, ordinary kriging and three Generalized Parameterization methods (that are combinations of previous zonation approaches) to parameterize hydraulic conductivity.		H	
Usunoff et al. (1992)	No	Three different models describing solute transport with the processes: i) Fickian dispersion and diffusion, ii) fickian dispersion and neglected diffusion and iii) non-fickian dispersion and diffusion.			H
Yakirevich et al. (2013)	Yes	Two models where one described a layered media and the other described a layered media with lenses based on boreholes.		C	
Ye et al. (2004)	No	Seven alternative variogram models for log permeability variations in unsaturated fractured tuff		H	
Ye et al. (2010), Reeves et al. (2010)	No	Five geological interpretations by three different companies. Three models are developed in response to non-unique interpretations of specific geological features (a thrust fault, a barrier to groundwater flow and a combination of the two). Five groundwater recharge scenarios informed by different methods (chloride mass balance, net infiltration method, Maxey-Eakin method) (I). Also included the effect of a surface water runon-runoff component and whether recharge occurs beneath a specific elevation in some models to test these hypothesis (H).	I/ H		I/H
Zeng et al. (2015)	No	Seven different representation of geometry by varying number of layers and the hydraulic conductivity distribution within the layers in a synthetic study.	H		

Study	Is the model objective well defined?	Conceptual multi-model development approach	Ph	Pr	SVS
Zhou and Herath (2016)	Yes	Three different models of geometry varying the number and extent of layers in a synthetic study.	H		
Zyvoloski et al. (2003)	Yes	To explain large hydraulic gradient a baseline model features a low permeability east-west zone, but there is no evidence for this feature, therefore three other models are proposed: i) Lower permeability hydrothermal alteration zone, ii) Alteration zone and NW-SE trending fault zone, iii) like the aforementioned but with additional fault features.	H		

A.2 MODEL TESTING LITERATURE REVIEW

Table A.2 Examples of approaches to test and make predictions with multiple plausible conceptual models. The 'Prior' column specifies if the prior probability in a Bayesian context is uninformed or informed by data or expert opinion. The sub-columns in the 'Model Testing' and 'Model Predictions' columns refer to modelling steps in the guideline by (Neuman and Wierenga, 2003). The fourth model testing step, the post-audit, is not included in this table as only one reviewed study (Nordqvist and Voss, 1996) applied this step. In the model testing steps the data type used for testing in the different steps are specified. In 'Model Prediction' the method used for ranking and making predictions is provided, where 'X' refers to methods not specified in the text. Additional data needs refers to the process of identifying additional data that could potentially discriminate between the conceptual models (as opposed to reducing parameter or prediction uncertainty).

Study	Prior	Model Testing			Model Predictions			
	Uninformed/ informed	Step 1	Step 2	Step 3	Model Ranking	Individual Predictions	Ensemble Predictions	Additional data needs
Altman et al. (1996)	-	-	-	Hydraulic conductivity.	-	X	-	-
Aphale and Tonjes (2017)	-	-	-	-	Area Metric	-	-	-
Carrera and Neuman (1986)	-	-	-	-	IC[1]	-	-	-
Castro and Goblet (2003)	-	-	-	Tracers	-	X	-	-
Elshall and Tsai (2014)	Informed	-	-	-	IC[1]	-	H-(ML)BMA[2]	-
Engelhardt et al. (2014)	-	-	-	Hydraulic conductivity	IC[1]	-	-	-
Feyen and Caers (2006)	Uninformed	Borehole data, seismic data, hydraulic conductivity.	-	-	-	-	X	-
Foglia et al. (2007)	-	-	-	-	IC[1], CV[3]	-	-	-
Foglia et al. (2013)	Uninformed	-	-	-	IC[1], X			

Study	Prior	Model Testing			Model Predictions			
	Uninformed/ informed	**Step 1**	**Step 2**	**Step 3**	**Model Ranking**	**Individual Predictions**	**Ensemble Predictions**	**Additional data needs**
Gedeon et al. (2013)	-	-	-	-	-	X	-	Sensitivity analysis
Harrar et al. (2003)	-	-	-	Transmissivity	-	X	-	-
He et al. (2014)	-	-	-	-	-	X	-	-
Hermans et al. (2015)	Uninformed	Geophysical data	-	-	-	-	-	-
Hills and Wierenga (1994)	-	-	-	Volumetric water content, solute concentrations	-	X	-	-
Højberg and Refsgaard (2005)	-	-	-	-	-	X	-	-
Johnson et al. (2002)	-	-	-	Drawdown	-	-	-	-
Kikuchi et al. (2015)	Uninformed	-	-	-	-	-	X	OD[4]
Knopman and Voss (1988)	-	-	-	-	-	X	-	OD[4]
Knopman and Voss (1989)								OD[4]
Knopman et al. (1991)								OD[4]
La Vigna et al. (2014)	-	-	Hydraulic head	-	-	-	-	-
Lee et al. (1992)	-	-	-	Tracer plume obs.	-	-	-	-
Li and Tsai (2009)	Uninformed	-	-	-	IC var[5]	-	MLBMA[6]	-
Linde et al. (2015)	-	Geophysical data	-	-	-	-	-	-
Lukjan et al. (2016)	Uninformed	-	-	-	IC[1]	X	-	-
Mechal et al. (2016)	-	-	-	Baseflow, transmissivity	IC[1]	X	-	-
Meyer et al. (2003)	Uninformed	-	-	-	IC[1]	-	MLBMA[6]	-
Meyer et al. (2007)	Uninformed	-	Hydraulic head, uranium concentrations	-	IC[1]	-	MLBMA[6]	-
Nettasana (2012)	Uninformed, informed	-	-	Hydraulic head	IC[1], GLUE[7]	-	GLUE-BMA[8], MLBMA[6]	-
Nettasana et al. (2012)	-	-	-	-	-	X	-	-
Nishikawa (1997)	-	-	-	Hydraulic conductivity.	-	X	-	-

Study	Prior	Model Testing			Model Predictions			
	Uninformed/ informed	**Step 1**	**Step 2**	**Step 3**	**Model Ranking**	**Individual Predictions**	**Ensemble Predictions**	**Additional data needs**
Nordqvist and Voss (1996)	-	-	-	-	-	X	-	OD[4]
Passadore et al. (2011)	-	Seismic data and stratigraphic records	-	-	-	X	-	-
Pham and Tsai (2015)	Uninformed	-	-	-	IC[1]	-	H-(ML)BMA[2]	OD[4]
Pham and Tsai (2016)	Uninformed	-	-	-	X	-	BMA[9]	OD[4]
Poeter and Anderson (2005)	-	-	-	Hydraulic conductivity. Model convergence.	IC[1]	-	X	-
Reeves et al. (2010)	Informed	-	-	-	X	-	X	-
Refsgaard et al. (2006)	-	-	-	-	-	X	-	-
Rogiers et al. (2014)	-	-	-	Hydraulic head	-	X	-	-
Rojas et al. (2008)	Uninformed	-	-	Hydraulic head, Model convergence.	-	-	GLUE-BMA[7]	-
Rojas et al. (2010a)	Uninformed	-	-	Hydraulic head	-	-	GLUE-BMA[7]	-
Rojas et al. (2010c)	Uninformed	-	-	Hydraulic head	IC[1]	-	MLBMA[6], AICMA, GLUE-BMA[7]	-
Samani et al. (2017)	Informed	-	-	Hydraulic head	IC[1]	-	-	-
Samper and Neuman (1989)	-	-	-	-	IC[1]	-	-	-
Schöniger et al. (2015)	Uninformed	-	-	Pumping tests	X	-	BMA[9]	-
Seifert et al. (2008)	-	-	-	Tritium apparent ages	-	X	-	-
Seifert et al. (2012)	-	-	-	Hydraulic conductivity	X	-	X	-
Selroos et al. (2002)	-	-	-	-	-	X	-	-
Troldborg et al. (2007)	-	-	-	CFC's, tritium and helium conc.	-	X	-	-

Study	**Prior**	**Model Testing**			**Model Predictions**			
	Uninformed/ informed	**Step 1**	**Step 2**	**Step 3**	**Model Ranking**	**Individual Predictions**	**Ensemble Predictions**	**Additional data needs**
Troldborg et al. (2010)	Uninformed	-	-	Hydraulic head, conductivity and TCE concentrations	-	-	BMA[9]	-
Tsai (2010)	Uninformed	-	-	-	IC var[5]	-	MLBMA[6]	-
Tsai and Elshall (2013)	Uninformed	-	-	-	IC var[5]	-	H-(ML)BMA[2]	-
Tsai and Li (2008)	Uninformed	-	-	-	IC var[5]	-	MLBMA[6]	-
Usunoff et al. (1992)	-	-	-	-	-	-	-	OD[4]
Yakirevich et al. (2013)	-	-	-	-	-	-	-	OD[4]
Ye et al. (2004)	Uninformed	-	-	-	IC[1], CV[3]	-	MLBMA[6]	-
Ye et al. (2010)	Informed	-	-	-	IC[1], GLUE[7]	-	GLUE-BMA[7]	
Zeng et al. (2015)	Uninformed	-	-	Hydraulic head? Model convergence.	-	-	GLUE-BMA[7]	-
Zhou and Herath (2016)	-	-	-	Water balance, travel time distribution.	IC[1]	-	-	-
Zyvoloski et al. (2003)	-	-	-	Flow paths are inferred from hydrogeochemical data	-	X	-	-

[1] Information Criteria including AIC, BIC, KIC etc. (IC)
[2] Hierarchal Bayesian Model Averaging (H-BMA)
[3] Cross-Validation (CV).
[4] Optimal design (OD).
[5] Information criterion corrected with variance window (IC var)
[6] Maximum Likelihood Bayesian Model Averaging (MLBMA)
[7] Generalized Likelihood Uncertainty Estimation Bayesian Model Averaging (GLUE-BMA).
[8] Generalized Likelihood Uncertainty Estimation (GLUE).
[9] Bayesian Model Averaging (BMA).

Appendix B: Chapter 3

The following describes the reasoning behind the prior ranges used for the stochastic water balance for Wildman River Area showed in Table 3.2. Since all prior ranges are based on the investigations by (Tickell and Zaar, 2017; Turnadge et al., 2018a), the reader is referred to these studies for further details on the study area.

B.1 RECHARGE

Net recharge has been estimated for the area using Chloride Mass Balance (CMB) method (Tickell and Zaar, 2017; Turnadge et al., 2018a) and environmental tracers (Turnadge et al., 2018a). The CMB method relies on the ratio of chloride concentration in local rainfall and in the groundwater. For the environmental tracers a lumped parameter model was used to identify an appropriate conceptual model and from that recharge rates were estimated in (Turnadge et al., 2018a) using closed-form solutions for age-depth and concentration-depth relationships as described in (Cook and Bohlke, 2000).

In (Tickell and Zaar, 2017) the net recharge was estimated based on the CMB method to 87 mm/y and 183 mm/y. In (Turnadge et al., 2018a) the net recharge was estimated to be between 32 mm/y and 178 mm/y for most of the study area based on the CMB method (Figure B.1).The estimates from the environmental tracer method agreed with this result. As most precipitation occurs in the wet season, it is valid to assume recharge in dry season is zero for the purpose of this water balance.

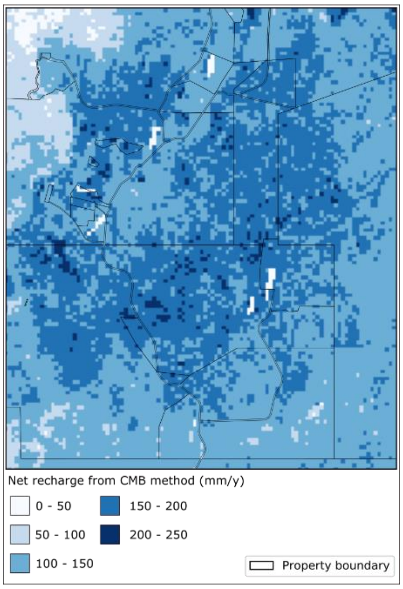

Figure B.1. Estimates of net recharge based on the CMB method (Turnadge et al., 2018a).

B.2 LATERAL OUTFLOW

Lateral outflow is thought to occur towards north following Swim Creek and northeast to Kakadu National park along respectively a northern and a southern palaeovalley. The northern palaeovalley has been assumed be connected northwards to thicker sequence known to occur towards the coast (Tickell and Zaar, 2017) (pp. 29), but very limited borehole data exists to support this assumption.

The lateral discharge is thought to consist of a component from each of the connected Mz/Cz sand aquifer and Koolpinyah Dolostone aquifer.

B.2.1 Transmissivity

In (Tickell and Zaar, 2017) a transmissivity value between 355 m^2/d and 2100 m^2/d based on pumping tests was used to estimate lateral discharge. A reinterpretation of pumping tests in (Turnadge et al., 2018a) (pp. 184) gave a range of transmissivity between 163 m^2/d and 1920 m^2/d for the 5th and 95th percentile for the sand aquifer based on 21 pumping tests. The transmissivity was estimated to be 109 m^2/d, 145 m^2/d, 295 m^2/d and 2630 m^2/d for the Koolpinyah Dolostone (Turnadge et al., 2018a) (pp. 140).

B.2.2 Width

The width of the sand aquifer can be constrained by borehole data (Figure B.2a), In the geological model developed in (Tickell and Zaar, 2017) (pp. 33) the width is 4 km and 6 km (for depths>10 m) for the southern and northern palaeovalley, respectively. However, the effective width of lateral outflow might be much more or much less. In our water balance the maximum width is constrained by the outcrops of impermeable rock on both sides of the palaeovalleys (7 km and 13 km for the southern and northern palaeovalley, respectively), while the minimum is set to 1 km for both boundaries.

The extent of the Koolpinyah Dolostone at the lateral boundaries can be constrained by observations of sinkholes and borehole observations. In (Tickell and Zaar, 2017) (pp. 22) the width is 1 km and 2.5 km for the north-eastern and northern boundary, respectively. The number of boreholes that include the Koolpinyah Dolostone is however very low (Figure B.2), and “sinkholes” are not known to be actually sinkholes developed on top of Dolostone.

In our water balance the minimum and maximum width is defined by making a concave to convex hull of the data points. The width of the northern boundary vary between 3 km and 10 km, while the north-eastern boundary vary between 1 km and 7 km.

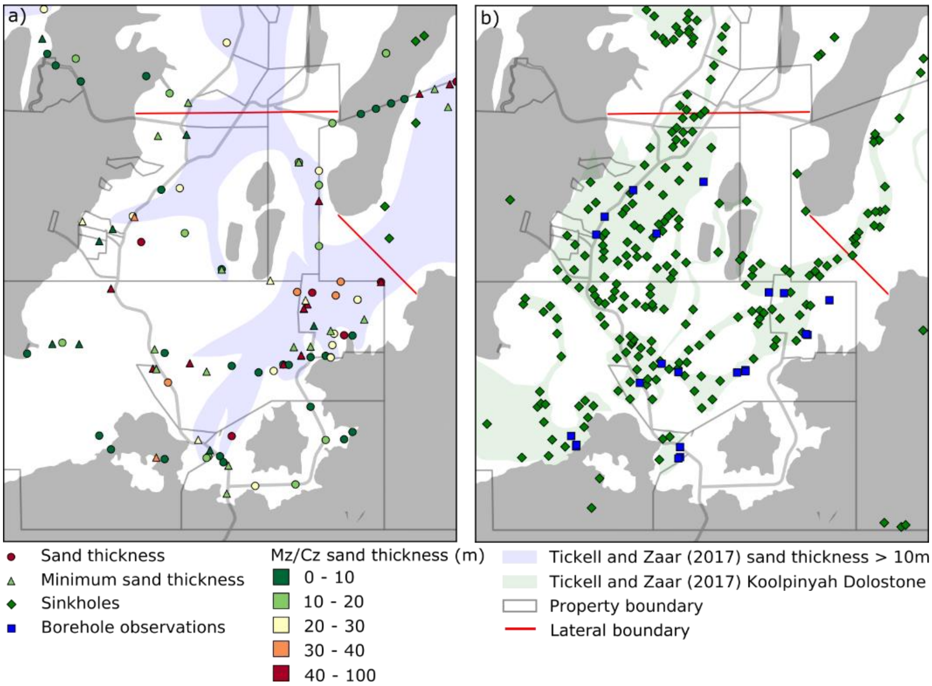

Figure B.2. (a) The observations of thickness of the Mz/Cz sand in boreholes, where triangle indicate boreholes where the bottom of the Mz/Cz sand has not been observed and the interpretation by (Tickell and Zaar, 2017). *(b) The location of observed sinkholes and boreholes observations of the Koolpinyah Dolostone with the interpretation of the extent of the Koolpinyah Dolostone by* (Tickell and Zaar, 2017).

B.2.3 Hydraulic gradient

Continuous observations of the water level was started immediately prior to the report by (Tickell and Zaar, 2017) and (Turnadge et al., 2018a) presents up to 9 months (August or November 2016 to May 2017) of hourly observation from 23 loggers (location seen in Figure B.3).

For the north-eastern boundary the loggers installed in the surficial leaky clay aquitard RN022961 and RN024174, and loggers installed in the semi-confined sand aquifer RN039073 and RN024667 can be used to constrain the gradient across the boundary (Figure B.3b). The gradient is respectively 0.001 and 0.0003 for the dry season, and 0.002 and 0.002 for the wet season between the two loggers. In our water balance the gradient is set to vary between 0.0001 and 0.001 for the dry season and 0.001 and 0.003 for the wet season.

Less data exist to constrain the gradient across the northern boundary. In our water balance the gradient is constrained by the head difference between RN024223 (Figure B.3b) and the ocean which is 0.0006 and 0.0008 in dry and wet season, respectively. For our water balance we set the gradient to vary between 0.0003 and 0.0009 for the dry season and between 0.0004 and 0.0012 for the wet season, respectively.

The gradient for the Dolostone aquifer is assumed to be similar as groundwater chemistry has revealed the two aquifers are hydraulically connected forming a regional aquifer (Tickell and Zaar, 2017) (pp. 42).

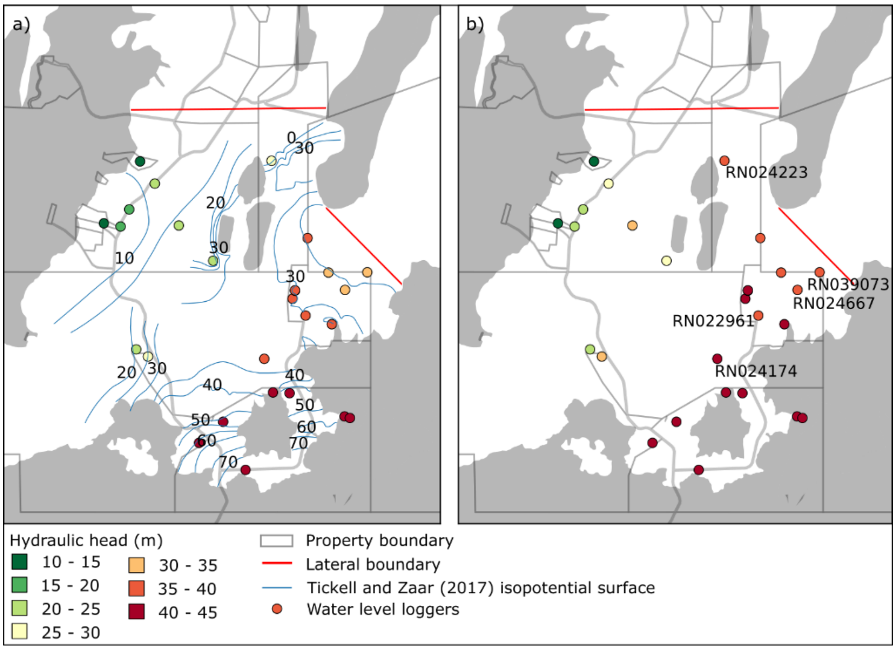

Figure B.3. Location of water level loggers and hydraulic head observation (a) at the end of the dry season (1. December 2016) and (b) at the end of the wet season (1. April 2017). (a) includes an interpretation of the isopotential surface by (Tickell and Zaar, 2017). *(b) includes the names of the water level loggers used for estimation of gradients across the lateral boundaries.*

B.3 BASEFLOW

B.3.1 Streams

Most surface water bodies in the area are ephemeral with flow resulting from rainfall-runoff (Turnadge et al., 2018a) (pp. 96). The exceptions are Jimmies and Opium Creek that are fed by springs. It is hypothesized that Swim Creek, Ben Bunga and Cattle Creek are also groundwater fed from diffuse discharge through the streambed.

In (Tickell and Zaar, 2017) the baseflow index was estimated for Jimmy's Creek and Swim Creek based on the mathematical recursive digital filter method developed by (Lyne and Hollick, 1979) of time series of streamflow. The baseflow index was estimated to 0.77, 0.7

and 0.65 for Opium Creek and 0.52, 0.46 and 0.3 for Swim Creek based on different baseflow separation techniques.

For the purpose of our water balance, baseflow is estimated stochastically for Jimmy's, Opium and Swim Creek streamflow observations with hydrograph separation. Jimmy's creek stream flow rate is not observed but estimated based on a relation with Opium Creek found in (Tickell and Zaar, 2017). We will use the hydrograph separation method described in (Eckhardt, 2005) to estimate the baseflow based on observed streamflow:

$$Q_{B,t} = \frac{(1-\beta)\cdot\alpha\cdot Q_{B_{t-1}} + (1-\alpha)\cdot\beta\cdot y_t}{1-\alpha\cdot\beta} \qquad \text{(B.1)}$$

Where the maximum baseflow index (β) represents the long-term ratio between baseflow and streamflow and the recession coefficient (α) represents the proportion of remaining streamflow on the next time step. y_t is the total streamflow at time step t and $Q_B$is the baseflow.The maximum baseflow index (β) was set to uniformly vary 0.25 and 0.6 for Swim Creek and 0.6 and 0.8 for Jimmy's and Opium Creek based on the results from the baseflow separation in Tickell and Zaar (2017). Eckhardt (2005) suggests a value of 0.8 for perennial streams and 0.5 for ephemeral streams with porous aquifers.

The recession coefficient (α) is set uniformly vary between 0.72 and 0.73 for Swim Creek and between 0.90 and 0.92 for Jimmy's Creek and Opium Creek. These numbers are based on the method specified in (Eckhardt, 2008), where stream flow at time step k is plotted against stream flow at time step k-1 for periods where streamflow is decreasing for five consecutive days. The slope of a linear regression that passes through the origin is the recession coefficient. A least squares regression has been applied and a 95 % confidence interval is used for the slope by assuming errors are normally distributed.

No continuous streamflow observations exist for Ben Bunga and Cattle Creek, but they are thought to behave similar to Swim Creek. The baseflow index for Swim Creek obtained from the hydrograph separation is therefore assumed to be representative for Ben Bunga and Cattle Creek. Point estimates of streamflow from (Tickell and Zaar, 2017) and (Turnadge et al., 2018a) are used to estimate baseflow.

B.3.2 Lagoons

Apart from streams and springs other surface water features that may be groundwater fed include Twin Sisters Lagoon, Number 1 Billabong, Lake Lucy and Mistake Billabong. Only the largest of the Twin Sisters lagoons has been subject to investigations.

In (Tickell and Zaar, 2017) an analysis of time series of water levels in the lagoon showed that at low water levels the rate of water level recession can be attributed to evaporation, while at high water levels there might be some net groundwater inflow to the lagoon. They estimated up to 12.5 % of the water balance in the dry season could not be accounted for by evaporation, which could correspond to a groundwater discharge of 0.75-1.8 mm/d. (Graham, 1985) also observed a higher evaporation rate than water level recession in the season 1984-85, which could be attributed to groundwater discharge at a rate of 2 mm/d. In lack of better data these values are extrapolated to other lagoons that are also thought to be depending on groundwater. We assume the same rate of groundwater discharge in the wet and in the dry season.

B.4 STORAGE

Turnadge et al. (2018a) provided an estimate of time required for the groundwater mounding in the wet season to dissipate. They estimated the time to 1 year which is consistent with the existing conceptualization as a fill-and-spill system (Tickell and Zaar, 2017; Turnadge et al., 2018a) (pp. 119). The annual storage ΔS_a is therefore thought to be around 0.

Appendix C: Chapter 4

C.1 DEPRESSIONS IN WILDMAN

In the period 25.-31. October 2018 in the end of the dry season, a fieldtrip was conducted in order to investigate the sinkholes in the Wildman River area. The objective was to investigate whether the sinkholes act as preferential recharge features. Five sinkholes were selected for focus of the investigation, spending one day at each sinkhole. These sinkholes were chosen based on their accessibility and vicinity to existing boreholes and water level loggers. It was prioritized to select sinkholes that were different regarding geometry, vegetation cover and from different parts of the area. It was not considered necessary to investigate sinkholes in both recharge and discharge area as they are thought to have the same geology regardless of hydraulic characteristics.

Prior to the fieldtrip a literature review established the location of depressions in the Wildman River Area. Table C.1 presents a short review of how the five investigated depressions have been mapped in literature. Tickell and Zaar (2017) mapped the location of doline features by indication location as point features based on satellite imagery and field visits. In a landform classification based on soil and vegetation samples as well as satellite imagery, the depressions were delineated as “flodded depressions and perennial billabongs” (Easey et al., 2016). Finally, the depressions have been outlined in the Water Observations from Space (WOfS) (Mueller et al., 2016), a dataset generated from Landsat-5 and Landsat-7, showing observations of surface water features over the period 1987-2014 in Australia with a resolution of 25 m. The orange polygons in Figure 4.2 delineate areas with at least a single occurrence of water in the WOfS dataset.

Table C.1. Short literature ***review of the five investigated sinkholes. Easey et al. (2016) present maps of the land resources in the area in the scale 1:25 000. Both are based on soil and vegetation samples as well as satellite imagery, respectively. Mueller et al. (2016) and present mapping of surface water bodies based on Landsat imagery in the scale of 25. Tickell and Zaar (2017) mapped sinkhole features by indication location as point features based on satellite imagery and field visits.***

Depression	Tickell and Zaar (2017)	Easey et al. (2016)	Mueller et al. (2016)
S1	Marked as sinkhole	11a (swamps, wetlands, flooded depressions and perennial billabongs)	Included
S2	Marked as sinkhole	Not in map extent	Included
S3	Marked as sinkhole	11a (swamps, wetlands, flooded depressions and perennial billabongs)	Included
S4	Marked as sinkhole	Not in map extent	Included
S5	Not included	11a (swamps, wetlands, flooded depressions and perennial billabongs)	Included

C.2 ALTERNATIVE MODEL DEVELOPMENT

C.2.1 Process structure models

The hydrogeological process structure of the depressions can be described in terms of their connection to the groundwater. When describing the depressions that containing seasonal or permanent water in the area, the terms lagoon, billabong, waterhole, pond, swamp and lake have been used interchangeably (Lloyd, 1999). We assume that the depressions could interact with groundwater in the same fashion as a lake.

Lakes interact with groundwater in three basic ways (Winter et al., 1998): 1) the lake receives groundwater discharge throughout the entire lakebed, 2) the lake losses groundwater recharge throughout the entire lakebed, and 3) part of the lake bed receives groundwater discharge while other parts of the lakebed losses groundwater recharge. Further, the depressions may not interact with groundwater at all. In case 2, the water level may be connected or disconnected from the water table, but we will not test the difference between the two. This result in four alternative understandings of the process structure (also see Figure 4.3). The water balance of the depression is influenced by:

- I: Evaporation only, there is no exchange between surface and groundwater

- II: Evaporation and groundwater recharge, where the groundwater flux occurs under either losing-connected or losing-disconnected conditions.
- III: Evaporation and groundwater discharge.
- IV: Evaporation, groundwater recharge and groundwater discharge.

C.2.2 Physical structure models

In developing alternative understandings of the physical structure beneath the depressions, we focus on understandings that would influence refraction seismic data and at the same time would affect whether the structure would allow preferential recharge.

Turnagde et al. (2018) suggested the depressions have formed as either dropout or buried sinkholes that develop on top of areas of relatively higher porosity where surface water infiltrates preferentially. The infiltrating water eventually results in dissolution of the dolomitic bedrock that forms underground cavities that eventually lead to collapse of the overlying features. This development history suggests that the permeability underneath the sinkhole is greater than outside the sinkholes.

A sinkhole developed in 2013 in the Darwin Region initially had steep walls, however over time the walls slumped and the sinkhole began to look like an older broad, shallow depression (Tickell, 2013). It has been hypothesized that over time a layer of low permeability clay and organic matter develop on top of the depressions sealing it from the groundwater (Graham, 1985; Schult and Welch, 2006).

The depth of investigation of the refraction seismic data is less than 30 m. Boreholes from the area show that the Dolostone is observed at depths between 40 and 100 m. The conceptual models are defined to have a maximum depth of 30 m and does therefore not include Dolostone.

Combination of the understandings of the development history of the depressions lead to four different physical structure models (also see Figure 4.3).

- A: Homogeneous sand subsurface without vertical stratification.
- B: High porosity/permeability zone promoting water infiltration.
- C: Homogenous sand with vertical stratification through a sealing clay layer.
- D: High porosity/permeability zone overlain by a sealing clay layer.

C.3 TESTING DATA

C.3.1 Process structure models (Y_{ps})

The remote sensing data consists of PlanetScope imagery (Planet Team, 2017) from 19 dates over the dry season; the first and last date respectively being 24-04-17 and 02-11-17. The PlanetScope satellites collects 4 bands (blue, green, red and near infrared) with a pixel size of 3 m. All the obtained imagery was captured during clear sky conditions.

Normalized Difference Water Index (NDWI) (McFeeters, 1996) can be used to assess the vegetation water content and the presence of open water bodies. NDWI compares how much near-infrared light (NIR) is reflected compared to visible green light (GREEN):

$$\mathrm{NDWI} = \frac{\mathrm{GREEN} - \mathrm{NIR}}{\mathrm{GREEN} + \mathrm{NIR}} \tag{C.1}$$

Here low values represent vegetation and soil features, while high values represent water bodies. Histograms of NDWI images that contain water will generally show a bimodal shape where the modes represent land and water pixels respectively.

The observation of interest is here the number of days after the start of the dry season, when the depressions run dry. The date where the depression is dry is interpreted through the application of a two-step approach to consider both spectral characteristics and spatial distribution of the water in the image:

1. Automatic thresholding. Segmentation into land and water based on thresholding.
2. Trajectory analysis. Correct all trajectories according to predefined rational rules.

Otsu’s method (Otsu, 1979) is used to perform automatic thresholding between water and non-water features when the NDWI image histogram is bimodal. The method searches for the optimal threshold that maximizes the inter-class variance by analysing the histogram of the image. The method is suitable for thresholding when the image histogram is bimodal but is less applicable when the water area is much smaller than the background. Rosin’s unimodal thresholding method (Rosin, 2001) is used when the NDWI image histogram is unimodal. The method searches for the point on a straight line between the histogram peak and the last empty bin that is furthest from the histogram. Following thresholding, the masks are cleaned by dilating with a one-pixel disk, filling donuts and eroded back again with the one-pixel disk.

A trajectory analysis (Powell et al., 2008; Zomlot et al., 2017) is performed that is based on the time series of each pixel to seek out classification errors. Classification errors can occur, e.g. because shallow water may not have a pure water signature, the view may be obstructed by tree canopy, or general noise may exist.

Since only images for the dry season have been obtained, we can assume the following rule applies in the trajectory analysis of the time series of each pixel: if a pixel is classified as wet and later is classified as dry, it is assumed to be correctly classified and considered to represent the drying out of the depression. For the remaining trajectories, the following rules are applied for correction (Figure C.1):

- Rule I: If a pixel has a trajectory that includes a sequence of dry-wet-dry, the pixel in the second time slice in this sequence is reclassified as dry.
- Rule II: If a pixel has a trajectory that includes a sequence of dry-wet-wet, the pixel in the first time slice in this sequence is reclassified as wet.
- Rule III: If a pixel has a trajectory that includes a sequence of dry-dry-dry, the rest of the pixels in the time slices are reclassified as dry.

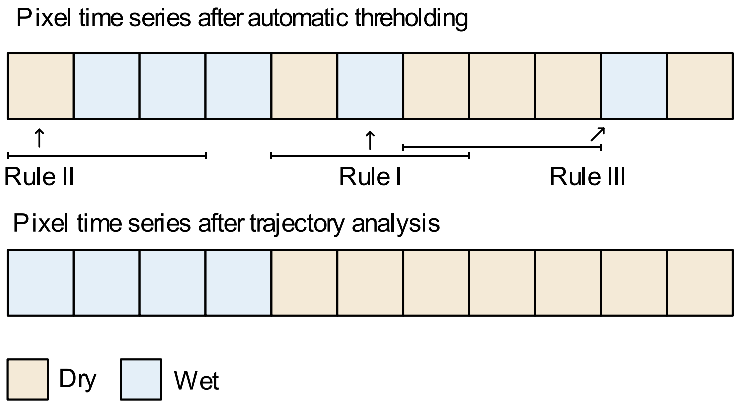
Pixel time series after automatic threholding
Rule II
Rule I
Rule III
Pixel time series after trajectory analysis
Dry
Wet

Figure C.1. Trajectory analysis application for correction of time series. Three rules (Rule I-III) are applied to correct trajectories extracted from automatic thresholding so that pixel time series can only go from wet to dry.

The likelihood function for the depressions being dry is defined as a beta distribution. Beta distributions are defined by four parameters: α, β, a lower limit and an upper limit. Solving the distribution for these four parameters requires the following assumption, which is based on the surface area of water over time for the depressions (Figure C.2):

- The lower limit is defined as the first date when less than 5 % of the maximum water-covered area is left in the depression.
- 10th percentile is defined as the first date when less than 3 % of the maximum area is left in the depression.
- 90th percentile is defined as the first date when less than 1 % of the maximum area is left in the depression.
- The upper limit is defined as the first date no water is observed in the depression.

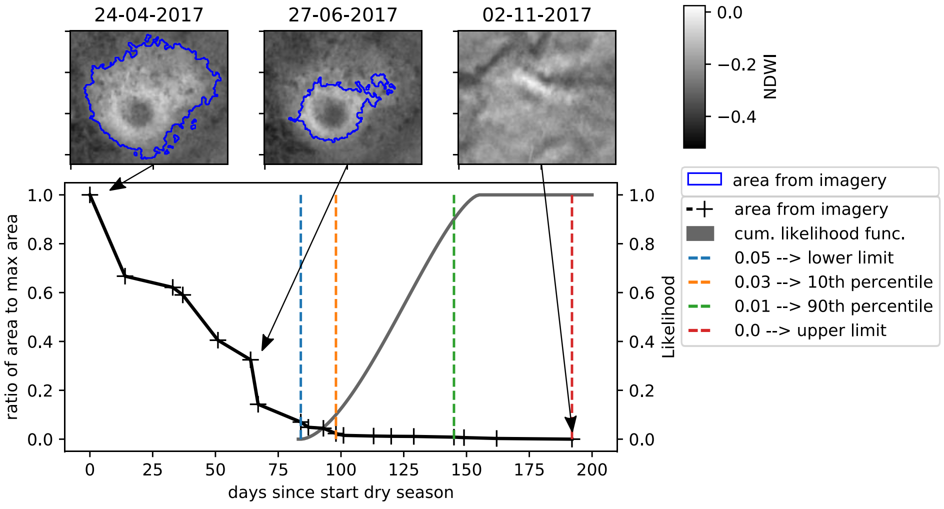

*Figure C.2. Likelihood **function definition based on timeseries of surface water area over time for depression S2.***

C.4 FORWARD MODELLING

C.4.1 Process structure models

A simple bucket water balance model was set up for each depression over the dry season (May-October) to calculate the day the depressions run dry:

$$\begin{gathered} inflow = outflow \pm \Delta d \\ d_t = d_{t-1} - inflow + outflow \end{gathered} \qquad \text{(C.2)}$$

where d_t is the water depth at timestep t and d_{t-1} is the water depth at timestep $t-1$. The result of the water balance model is the timestep, where $d_t = 0$. The inflow will consist of groundwater discharge (zero rainfall inflow during the dry season) and the outflow of groundwater recharge and evaporation according to the applied process structure.

The prior parameter probabilities used in the bucket water balance for all depressions are defined in Section C.5.

C.4.2 Physical structure models

In order to compute the travel-times from a known subsurface structure we need to define relationships between the elastic properties and how these vary as a function of porosity and water saturation. Instead of picking a velocity range for each facies in our physical structure models we elected to utilize rock physics relationships to provide reasonable estimates for the velocities expected by our structure. This section outlines the rock physics relationships we used to calculate velocities given a specific lithology (e.g. clay or sand) and a known porosities and water saturation. From these parameters our relationships provide a p-wave velocity that can be fed into a program to calculate the first-arrival travel times. The first arrival travel-times can be directly compared to ones measured in the field. The well-known relationship between bulk modulus, shear modulus and density and velocity are given (Telford and Telford, 1976):

$$V_p = \sqrt{\frac{K + \frac{4}{3}G}{\rho}} \tag{C.3}$$

$$V_s = \sqrt{\frac{G}{\rho}} \tag{C.4}$$

$$\rho = \rho_m(1-\phi) + (s_w\rho_w + (1-s_w)\rho_a)\phi \tag{C.5}$$

where V_p is p-wave velocity in km/s, V is the s-wave velocity in km/s, K is the bulk modulus in GPa, G is the shear modulus in GPa, ρ is the density in g/cm^3, ϕ is porosity and s_w is the water saturation. Here we calculate density as a function of porosity and water saturation. The density is calculated prior to the velocity calculation as a linear average. ρ_m represents the density of the solid phase, ρ_w the density of water, and ρ_a the density of air.

To calculate the elastic properties at different porosities and water saturations we use a porous rock physics model based on Hertz–Mindlin contact theory (Helgerud, 2001; Helgerud et al., 1999; Mindlin, 1949) and modified Hashin-Shtrikman bounds (Hashin and Shtrikman, 1963). The step to calculate the elastic properties follows three steps. First, we must calculate the elastic moduli at the critical porosity (Eq. (C.7). The critical porosity is defined as the porosity at which the rock goes from grain supported to fluid supported (Nur et al., 1998). Second, we define the elastic properties between zero and the critical porosity using a Hashin-Shtrikman relationship (Eq. (C.8 and(C.9) (Hashin and Shtrikman, 1963). Third, to ensure that we define a velocity for any given input we define the elastic moduli above the critical porosity using a Reuss average (Eq. (C.10) (Dvorkin et al., 1999; Nur et al., 1998). The critical porosity is defined as the porosity at which the rock goes from being grain supported to being fluid supported (Nur et al., 1998). After we have obtained the elastic properties at different porosities, we use a two more relationships to account for influence of fluid on velocity. First, we estimate the bulk modulus of the fluid air mixture in the pores (Eq. (C.12).

Second, we use Gassmann's equation (Eq. (C.11) to predict the saturated bulk modulus. Finally, the bulk and shear modules are converted to velocity (Eq. (C.3 and (C.4).

The Hertz-Mindlin model, we calculate the elastic properties at a critical porosity and effective pressure, or depth, using Eq. (C.7).

$$K_{HM} = \sqrt[3]{\frac{C^2(1-\phi_c)^2 G_m^2}{18\pi^2(1-\upsilon)^2} P} \tag{C.6}$$

$$G_{HM} = \frac{5-4\upsilon}{5(2-\upsilon)} \sqrt[3]{\frac{3C^2(1-\phi_c)^2 G_m^2}{2\pi^2(1-\upsilon)^2} P} \tag{C.7}$$

where K_{HM} and G_{HM} are the bulk and shear moduli at the critical porosity (ϕ_c), P is the effective pressure in GPa, υ is Possion's ratio of the material, and C is the number of contacts per grain. Poisson's ratio is calculated using the bulk and shear moduli defined by literature (Table C.2). The critical porosity is approximately 0.4 for sandstones but can be as low as 0.05 for igneous rocks (Mavko et al., 2009; Nur et al., 1998).

To find the effective moduli of the dry rock (K_{eff} and G_{eff} f) at porosities between zero and the critical porosity, we use the modified Hashin-Shtrikman (HS) lower bound (Eq. (C.8) and upper bound (Eq. (C.9). We use both of these boundaries to constrain possible elastic moduli (Mavko et al., 2009). The upper Hashin-Shtrikman boundary is often time referred to as the "stiff" sand model and the lower boundary is referred to as the "soft" sand boundary. The equations for the lower bounds are the following:

$$K_{eff(\phi<\phi_c)} = \left(\frac{\phi/\phi_c}{K_{HM}+\frac{4}{3}G_{HM}} + \frac{1-\phi/\phi_c}{K_m+\frac{4}{3}G_{HM}}\right)^{-1} - \frac{4}{3}G_{HM}$$

$$G_{eff(\phi<\phi_c)} = \left(\frac{\phi/\phi_c}{G_{HM}+\frac{G_{HM}}{6}\left(\frac{9K_{HM}+8G_{HM}}{K_{HM}+2G_{HM}}\right)} + \frac{1-\phi/\phi_c}{G_m+\frac{G_{HM}}{6}\left(\frac{9K_{HM}+8G_{HM}}{K_{HM}+2G_{HM}}\right)}\right)^{-1} - \frac{G_{HM}}{6}\left(\frac{9K_{HM}+8G_{HM}}{K_{HM}+2G_{HM}}\right) \quad \text{(C.8)}$$

$$K_{eff(\phi<\phi_c)} = \left(\frac{\phi/\phi_c}{K_{HM}+\frac{4}{3}G_m} + \frac{1-\phi/\phi_c}{K_m+\frac{4}{3}G_m}\right)^{-1} - \frac{4}{3}G_m$$

$$G_{eff(\phi<\phi_c)} = \left(\frac{\phi/\phi_c}{G_{HM}+\frac{G_m}{6}\left(\frac{9K_m+8G_m}{K_m+2G_m}\right)} + \frac{1-\phi/\phi_c}{G_m+\frac{G_m}{6}\left(\frac{9K_m+8G_m}{K_m+2G_m}\right)}\right)^{-1} - \frac{G_m}{6}\left(\frac{9K_m+8G_m}{K_m+2G_m}\right) \quad \text{(C.9)}$$

For porosities higher than the critical porosity, we use Reuss average (Dvorkin et al., 1999; Nur et al., 1998) (Eq. (C.10):

$$K_{eff(\phi>\phi_c)} = \left(\frac{(1-\phi)/(1-\phi_c)}{K_{HM}+\frac{4}{3}G_{HM}} + \frac{(\phi-\phi_c)/(1-\phi_c)}{\frac{4}{3}G_{HM}}\right)^{-1} - \frac{4}{3}G_{HM}$$

$$G_{eff(\phi>\phi_c)} = \left(\frac{(1-\phi)/(1-\phi_c)}{G_{HM}+\frac{G_m}{6}\left(\frac{9K_m+8G_m}{K_m+2G_m}\right)} + \frac{(\phi-\phi_c)/(1-\phi_c)}{G_m+\frac{G_m}{6}\left(\frac{9K_m+8G_m}{K_m+2G_m}\right)}\right)^{-1} - \frac{G_m}{6}\left(\frac{9K_m+8G_m}{K_m+2G_m}\right) \quad \text{(C.10)}$$

The last step is to include the fluid effect on elastic properties for various saturations. In near-surface applications, the saturating fluid will be a mixture of water and air. A common way to model how saturation influences the elasticity is through an effective fluid model (Mavko et al., 2009). Once an effective fluid bulk modulus K_{fl} is calculated, it can be included in the rock physics model using into Gassmann’s equations (Mavko et al., 2009) to calculate the saturated-rock bulk modulus K_{sat} (Eq. (C.11). The shear modulus G_{sat} is not influenced by saturation, so it is equal to the effective shear modulus G_{eff} of the dry rock.

$$\frac{K_{sat}}{K_m - K_{sat}} = \frac{K_{eff}}{K_m - K_{eff}} + \frac{K_{fl}}{\phi(K_m - K_{fl})} \qquad \text{(C.11)}$$

$$G_{sat} = G_{eff}$$

To calculate the effective fluid bulk modulus (K_{fl}), we use Brie’s equation (Brie et al., 1995) (Eq. (C.12).

$$K_{fl} = (K_w - K_a)(1 - s_a)^e + K_a \qquad \text{(C.12)}$$

where K_w is the bulk modulus of water (2.1 GPa) and K_a is the bulk modulus of air (0.01 GPa), s_a is the air saturation, and e is an empirical parameter usually equal to 3 (Mavko et al., 2009).

Using the rock physics relationships (Eq. (C.3-(C.12) we can now define the physical structure in terms of velocity. This is a critical step because the forward model code requires velocity values, not lithology, porosity, or density. Given the conceptualization and the rock physics modelling we can now produce a saturated and unsaturated velocity for the clay, the sand anon the high porosity sand. From this step, the velocities that represent each physical structure (Figure C.3) can be easily projected onto a triangular mesh and fed to pyGIMLi (Rücker et al., 2017) to compute the travel-times. Using the same configuration that we collected the data within the field, that is 36 geophones spaced equally at 4 m with a shot every four meter we can calculate the travel time from each source to each receiver. The ray tracing algorithm in pyGIMLi is based on a shortest path algorithm (Moser, 1991). Since there is not inversion involved the forward modelling process runs quickly making it possible to run many realizations and produce travel-time curves for 1000s of models. The prior parameter probabilities used in the refraction simulation and the rock physics model are defined in Section C.5.

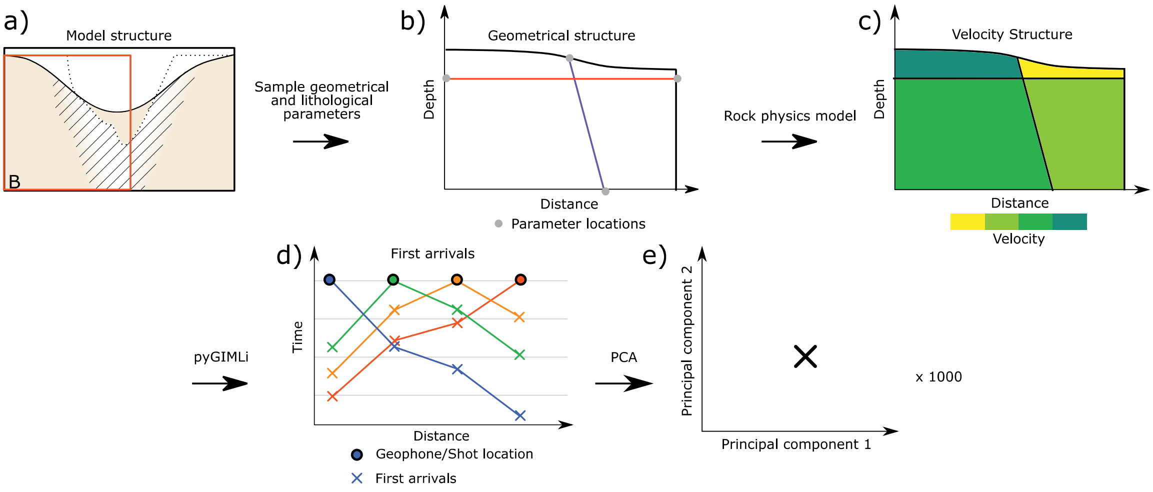

Figure C.3. Workflow for generating one realization of the physical model structure. One of four model structures (Figure 4.3) is chosen (a) and geometrical parameters are sampled to translate the model structure into a geometrical structure (b) (Figure C.5). Further lithological parameters are sampled as input to a rock physics model that translate the lithologies into velocities (c). The velocity structure is used as input to pyGimli to obtain a realization of first arrivals (d). The dimensions of the realization of first arrivals is reduced in a Principal Component Analysis (PCA) to two principal components (e). This workflow is repeated 1000 times for each realization.

C.5 ON DEFINING PRIOR RANGES FOR PARAMETERS

The following describes the reasoning behind the prior probability ranges for the parameters used in the forward geophysical model and bucket water balance model for the depressions. All prior parameter ranges are described by uniform probability distributions defined by the minimum and maximum value, so that all parameter values in the ranges are equally likely.

Some parameters depend on the geometry of the depressions and a few geometrical properties are therefore defined in the following. The *outline* of the depression is here defined as the maximum water filled area based on the PlanetScope imagery in the start of the dry season (24/4-2017). The *edge* of the sinkhole is defined as the point on the seismic line that crosses the outline and the *centre* is defined as the point on the seismic line furthest from the outline within the depression. This, with exception of depression S4 where the seismic line is too short to cross near the geometrical centre of the depression because of its size. Here the centre is defined as the point in the depression furthest

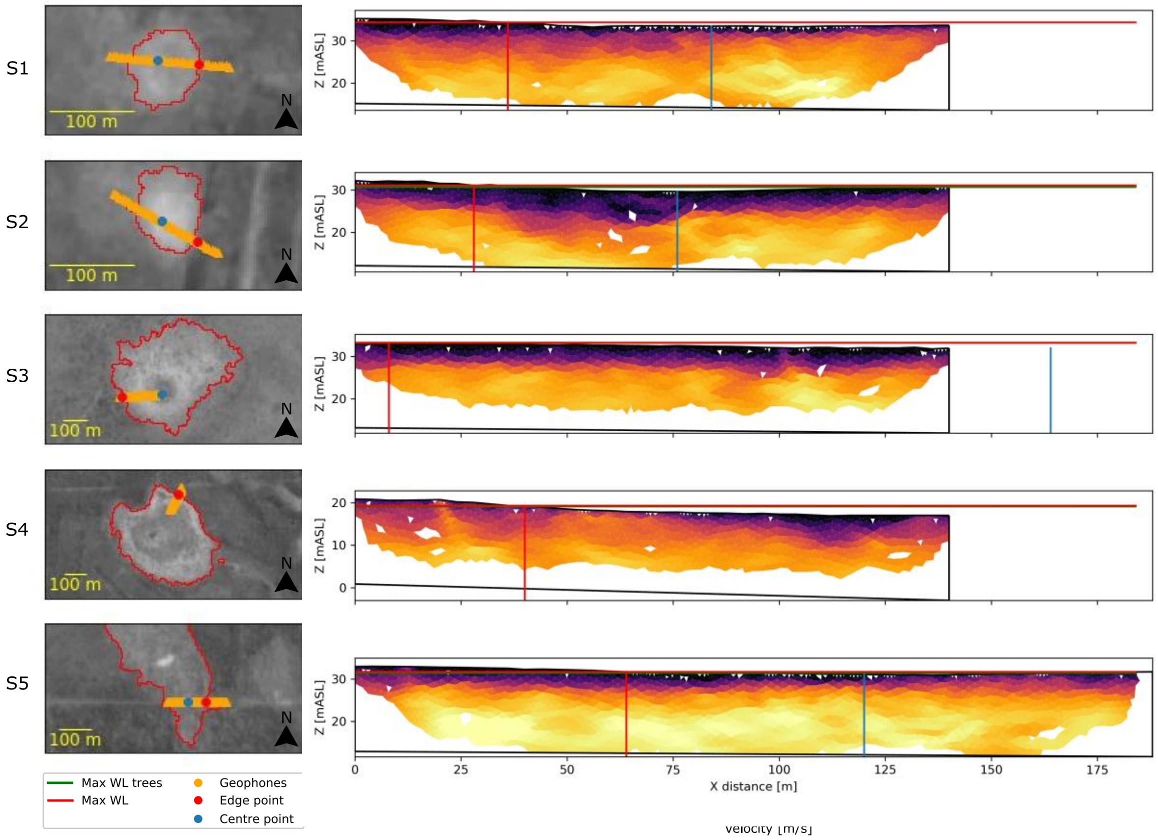

Figure C.4. Geometry and seismic inversions of the seismic refraction data of the five depressions (S1 to S5).

C.5.1 Physical structure

The prior probabilities of the parameters for the forward geophysical model is shown in Table C.2. The parameters in the geophysical model can be divided into geometrical parameters that control the zonation of the cross-section and lithological parameters that determine the velocity within the different zones used in rock physics model.

Table C.2. Prior parameter probabilities for the physical structure for each depression (S1 to S5).

	S1	S2	S3	S4	S5	Unit
W.Z.O	U(30.4, 32.4)	U(27.1, 29.1)	U(29.1, 31.1)	U(11.6, 16.8)	U(18, 29)	m
W.Z.I	U(30.4, 32.4)	U(27.1, 29.1)	U(29.1, 31.1)	U(11.6, 16.8)	U(18, 29)	m
W.X	U(36, 79)	U(28, 71)	U(8, 135)	U(40, 135*)	U(64, 115)	m
P.X.U	U(36, 79)	U(28, 71)	U(8, 135)	U(40, 135*)	U(64, 115)	m
P.X.L	U(36, 79)	U(28, 71)	U(8, 135)	U(40, 135*)	U(64, 115)	m
C.X	U(0, 79)	U(0, 71)	U(0, 135)	U(0, 135*)	U(0, 115)	m
C.Z	U(28.5, 32.5)	U(24.6, 28.6)	U(27, 31)	U(11.8, 15.8)	U(26, 30)	m
θ_h	U(θ_t + 0.05, θ_t + 0.15)					-
θ_t	U(0.2, 0.4)					-
K_{clay}	U(5, 12)					GPa
K_{sand}	U(12, 18)					GPa
G_{clay}	U(3, 7)					GPa
G_{sand}	U(5, 11)					GPa
ρ_{clay}	U(2.6, 2.75)					g/cm^3
ρ_{sand}	U(2.6, 2.65)					g/cm^3

C.5.1.1 Geometrical priors

The geometry in the sinkholes are described in the cross-section where the seismic data has been collected. The structure of the model is described by lines that are defined by points, that make up the geometrical parameters (Figure C.5). The lines intersect and create polygons or zones of similar velocity.

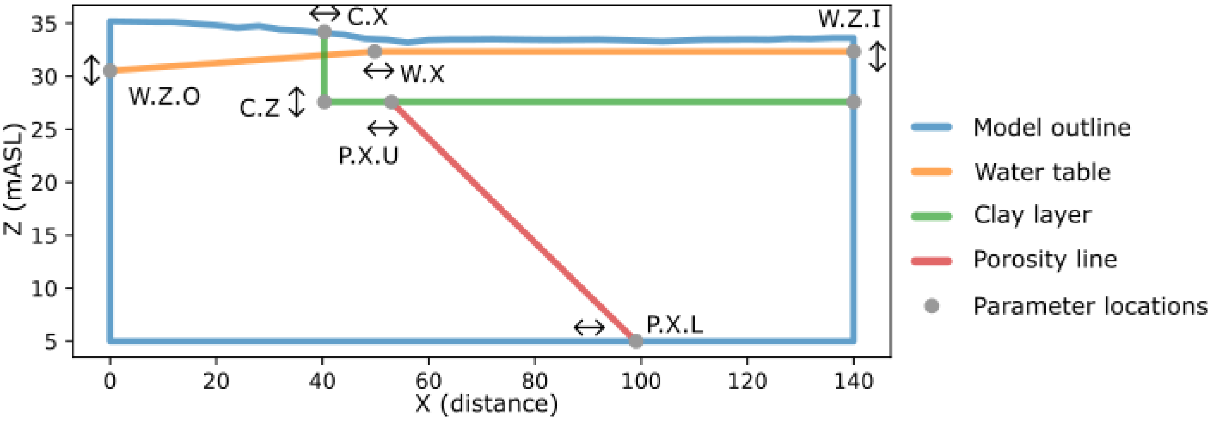

Figure C.5. Geometrical parameters ***for the geophysical forward model. The parameters described with a X varies horizontally, while parameter names including a Z varies vertically.***

Water table

The water table line describes the location of the water table at the time of the seismic survey. The water table is defined by three points. In some of the sinkholes, auger holes were drilled deep enough to meet the water table, while in the other sinkholes, the definition of the priors must rely on water table loggers in the vicinity of the sinkholes.

The water table logger RN24177, 20 m away from S1, measured 31.4 mASL at the time of the survey. In the model the water table is set to vary between 30.4 mASL and 32.4 mASL for both the water table z coordinates (W.Z.O and W.Z.I). In S2 the water table was measured in an auger hole in the middle of the sinkhole at 28.1 mASL at the time of the survey. In the model the water table is set to vary between 27.1 mASL and 29.1 mASL. In S3, the water table was measured in an auger hole in the middle of the sinkhole at 30.1 mASL at the time of the survey. In the model the water table is set to vary between 29.1 mASL and 31.1 mASL. S4 is located between logger RN039077 in north-north-west (1.5 km away) and logger RN039769 in east-south-east (6 km away). The water level at the time of the survey was 11.6 mASL and 18.3 mASL at logger RN039077 and RN039769, respectively. An auger hole of 1 m depth (16.9 mASL) in the middle of the sinkhole did not meet the water table and the prior is therefore set to vary between 11.6 mASL and 16.8 mASL. S5 is located between logger RN039769 in east-north-east (3 km away) and logger RN038022 in north-north-east (3 km away). The water level at the time of the survey was 18.28 mASL and 18.82 mASL at logger RN039769 and RN038022, respectively. A 1.9 m deep (29.1 mASL) auger hole did not meet the water table. The water level is set to vary between 18 mASL and 29 mASL at S5.

The prior range for the parameter controlling the location of the bend in the water table (W.X) is estimated based on topography of the seismic cross-section. The lower limit is set at the edge of the sinkhole while the upper limit is set 5 m from the centre of the depressions.

Porosity line

The porosity describes a sharp contrast in porosity between the lithology within and outside the sinkholes. It is described by two parameters, P.X.U and P.X.L, that vary horizontally and are defined within the same prior range. The lower limit is set at the edge of the sinkhole while the upper limit is set 5 m from the centre of the sinkhole.

Clay line

The clay line describes a sharp boundary between sand and clay in the sinkhole. The clay layer represents the sediments that are deposited on top of the sinkhole after the collapse. From other sinkholes in the area, it is known sinkholes are not more than 5 m deep (Schult and Welch, 2006). The upper bound for the clay layer is therefore set to 5 m below the elevation in the centre. The lower bound is set to 1 m below the elevation in the centre.

The C.X variable controls the lateral extent of the clay layer. The lower limit is set at the edge of the profile while the upper limit is set 5 m from the centre of the sinkhole.

C.5.1.2 Lithological property priors

The velocity in the different zones described in the geometrical model is determined by a rock physics model.

The parameters for this model include porosity, bulk modulus, shear modulus and the density for which a prior probability distribution will be defined. It also includes the water saturation, the Hertz-Mindlin grain contact parameter and the critical porosity which are all kept constant. Generally, these values have not been measured in the area and must be estimated from literature.

In each realisation the rock physics model is applied for each velocity zone. Each input value drawn from the uniform distribution (Table C.2) is used as the mean value from which a normal distribution is defined. The values in each mesh grid cell is populated from this normal distribution. The standard deviation of the normal distribution is defined as (max-min)/10.

Water saturation

The water saturation is a constant value of 0.2 for unsaturated zone and 0.99 for the saturated zone.

Grain contact parameter

The grain contact parameter is the average number of contacts per grain which generally ranges between 4 and 12 (Mavko et al., 2009). We apply a constant value of 4 for the grain contact parameter as an initial sensitivity analysis showed that this parameter is relatively insensitive.

Critical porosity

The critical porosity is the porosity where rock goes from being grain supported to being fluid supported. A constant value of 0.4 is used as the critical porosity, which is the critical porosity of sandstone (Mavko et al., 2009). As for the grain contact parameter an initial sensitivity analysis showed that this parameter is relatively insensitive.

Porosity ($\boldsymbol{\theta_h}, \boldsymbol{\theta_t}$)

Manger (1963) summarised 900 data items of porosity for sedimentary rock mostly from America, Great Britain, Germany and Switzerland. According to Manger (1963) the porosity for both sand and clay can range between 0.2 and 0.5. We are interested in defining a higher porosity θ_h zone in the middle of the depressions and a zone of more typical values of porosity θ_t around the depression. The latter zone is set to vary between 0.2 and 0.4, while the high porosity zone is set to always have a porosity between 0.05 and 0.15 higher than the typical porosity zone.

Bulk modulus ($\boldsymbol{K_{clay}}, \boldsymbol{K_{sand}}$)

Literature values for shear modulus has been found in (Vanorio et al., 2003; Wang et al., 2001). The bulk modulus has been reported to have values for sand of 13-17 GPa and for clay of 6-11 GPa. In the forward model for the seismic cross-section the bulk modulus is set to vary between 12 GPa and 18 GPa for sand and between 5 GPa and 11 GPa for clay.

Shear modulus ($\boldsymbol{G_{clay}}, \boldsymbol{G_{sand}}$)

Literature values for shear modulus has been found in (Vanorio et al., 2003; Wang et al., 2001). The shear modulus has been reported to have values for sand of 6-10 GPa and for clay of 4-6 GPa. In the forward model for the seismic cross-section the shear modulus is set to vary between 5 GPa and 11 GPa for sand and between 3 GPa and 7 GPa for clay.

Mineral density ($\boldsymbol{\rho_{clay}}, \boldsymbol{\rho_{sand}}$)

In an x-ray diffraction analysis of sediment samples from Wildman River Area, Turnadge et al. (2018) found that the sand mineralogical composition is mainly quartz with minor fraction of kaolin. The sandy clay is also mainly composed of quartz (~65 %) but contain more kaolin (~30 %). Further, the sandy clay contains minor fractions of muscovite and goethite. The mineral density of quartz and kaolin is 2.6-2.65 g/cm3 while it is 3.3-4.3 g/cm3 for goethite and 2.77-2.88 g/cm3 ("Mineralogy Database," n.d.). We set the prior for sand density to vary between 2.6 g/cm3 and 2.65 g/cm3 and for clay between 2.6 g/cm3 and 2.75 g/cm3 considering the mineralogy.

C.5.2 Process structure

The prior ranges for the parameters in bucket water balance based on the alternative understandings of the process structures of the depressions is presented in Table C.**3**.

Table C.3. Prior parameter probabilities ***for the process structure models.***

	S1	S2	S3	S4	S5	Unit
d_{max}	U(1.2, 1.6)	U(1.2, 1.6)	U(1.7, 1.9)	U(2.3, 2.6)	U(1.3, 2)	m
E	U(5, 9)					mm/d
q_{rec}	U(0.5, 2.5)					mm/d
q_{dis}	U(0.5, 2.5)					mm/d

Maximum water depth ($\boldsymbol{d_{max}}$)

The five depressions have not been visited in the end of the wet season and the maximum water depth have therefore not been observed. From trees in depressions however, high water marks (Figure C.6) could be identified all depressions except S1. These values can be compared with values obtained by comparing maximum water filled area of the sinkholes (24/4-17) with elevations obtained in the field, see Figure C.4.

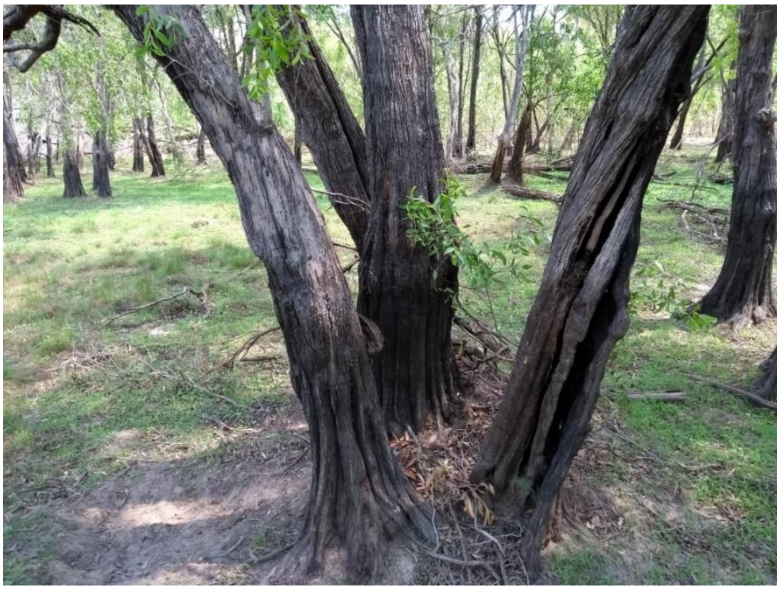

Figure C.6. High water ***level marks on trees in sinkhole 153 used to estimate maximum water depth in wet season.***

Pan evaporation (E)

Tickell and Zaar (2017) used pan evaporation in the dry season ranging from 5.9 to 7.9 mm/day. Rayner (2005) presented a modelled Class A pan evaporation dataset for Australia based on a simple linear combination of gridded solar radiation and vapour pressure deficit. The pan evaporation is about 5 – 9 mm/day from May to September in the Top End in this dataset. This range is used as prior range in the water balance for the depressions.

Groundwater recharge and discharge ($\boldsymbol{q_{rec}}$, $\boldsymbol{q_{dis}}$)

The groundwater recharge/discharge have not been investigated for the five depressions. However, the major Twin Sisters Lagoon, that is assumed to have same development history, has been the subject of to some investigation.

In an analysis of time series of water levels in the lagoon, Tickell and Zaar (2017) showed that at low water levels the rate of water level recession can be attributed to evaporation, while at high water levels there might be some net groundwater inflow to the lagoon. They estimated up to 12.5% of the water balance in the dry season could not be accounted for by evaporation, which could correspond to a groundwater discharge of 0.75–1.8 mm/day. Graham (1985) also observed a higher evaporation rate than water level recession in the season 1984–85, which could be attributed to groundwater discharge at a rate of 2 mm/day. In lack of better data these values used as prior parameter range for groundwater recharge/discharge in the water balances.

Appendix D: Publication list

D.1 PUBLICATIONS

Enemark, T., Peeters, L. J. M., Mallants, D., & Batelaan, O. Using data for model development in a multi-model framework. Manuscript in preparation.

Enemark, T., Peeters, L. J. M., Mallants, D., Flinchum, B. A. & Batelaan, O., in Review. A systematic approach to hydrogeological conceptual model testing combining remote sensing and geophysical data.

Enemark, T., Peeters, L. J. M., Mallants, D., & Batelaan, O. (2019). Hydrogeological conceptual model building and testing: A review. Journal of Hydrology, 569. https://doi.org/10.1016/j.jhydrol.2018.12.007

Enemark, T., Peeters, L. J. M., Mallants, D., Batelaan, O., Valentine, A. P., & Sambridge, M. (2019). Hydrogeological Bayesian Hypothesis Testing through Trans-Dimensional Sampling of a Stochastic. Water, 11(1463). https://doi.org/10.3390/w11071463

D.2 CONFERENCE ABSTRACTS

Enemark, T., Peeters, L. J. M., Mallants, D., & Batelaan, O. (2020). Systematic hydrogeological conceptual model testing using remote sensing and geophysical data. *EGU General Assembly 2020, Vienna, Austria, 3-8 May 2020. Copernicus / European Geosciences Union (EGU).* Retrieved from https://doi.org/10.5194/egusphere-egu2020-11688.

Enemark, T., Peeters, L. J. M., Mallants, D., & Batelaan, O. (2019). Hydrogeological Bayesian hypothesis testing through trans-dimensional sampling of a stochastic water balance model. *Australasian Groundwater Conference 2019, Brisbane, Australia, 24-27 November 2019. NCGRT / IAH. Abstract 460.* Retrieved from www.groundwater.com.au/documents/agc2019-book-of-abstracts-updated.pdf

Enemark, T., Peeters, L. J. M., Mallants, D., & Batelaan, O. (2019). Characterisation of conceptual uncertainty through alternative model development. *EGU General Assembly 2019, Vienna, Austria, 7-12 April 2019. Copernicus / European Geosciences Union (EGU).* Retrieved from https://ui.adsabs.harvard.edu/abs/2019EGUGA..21.4547E/abstract

Enemark, T., Peeters, L. J. M., Mallants, D., & Batelaan, O. (2018). Addressing conceptual uncertainty through a multi-model approach using bold hypotheses. *Australian Geoscience Council Convention (AGCC), Adelaide, Australia 14-18 October 2018. Australian Geoscience Council. Abstract 526.* Retrieved from https://www.agcc.org.au/wp-content/uploads/2019/03/clickheretodownloadabstract.pdf

Enemark, T., Peeters, L. J. M., Mallants, D., & Batelaan, O. (2018). The use of bold hypotheses to systematically develop hydrogeological multi-model conceptualizations. *IAH 2018 - 45th Congress of The International Association of Hydrogeologists; 9-14 September 2018; Daejeon, Korea. International Association of Hydrogeologists (IAH); 2018. p.67.* http://hdl.handle.net/102.100.100/86513?index=1

Enemark, T., Engesgaard, P. (2017). Does resolution in surface water body representation in groundwater models affect nitrate transport and removal capacity in a coupled groundwater-surface water system? *Australasian Groundwater Conference 2017, Sydney, July 11-13. International Association of Hydrogeologists.* Retrieved from http://hdl.handle.net/102.100.100/88302?index=1

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle, in: Petrov, B.N., Csaki, F. (Eds.), Second International Symposium on Information Theory. Budapest, pp. 261–304.

Altman, S.J., Arnold, B.W., Barnard, R.W., Barr, G.E., Ho, C.K., McKenna, S.A., Eaton, R.R., 1996. Flow Calculations for Yucca Mountain Groundwater Travel Time (GWTT-95). Report SAND96-0819. Albuquerque, New Mexico, USA.

Anderson, M.P., Woessner, W.W., 1992. The role of the postaudit in model validation. Adv. Water Resour. 15, 167–173. https://doi.org/10.1016/0309-1708(92)90021-S

Anderson, M.P., Woessner, W.W., Hunt, R.J., 2015a. Modeling Purpose and Conceptual Model, in: Anderson, M.P., Woessner, W.W., Hunt, R.J. (Eds.), Applied Groundwater Modeling. Elsevier Inc, pp. 27–67. https://doi.org/http://dx.doi.org/10.1016/B978-0-08-091638-5.00002-X

Anderson, M.P., Woessner, W.W., Hunt, R.J., 2015b. Basic Mathematics and the Computer Code, in: Anderson, M.P., Woessner, W.W., Hunt, R.J. (Eds.), Applied Groundwater Modeling. Elsevier Inc, pp. 69–114. https://doi.org/https://doi.org/10.1016/B978-0-08-091638-5.00003-1

Aphale, O., Tonjes, D.J., 2017. Multimodel Validity Assessment of Groundwater Flow Simulation Models Using Area Metric Approach. Groundwater 55, 219–226. https://doi.org/10.1111/gwat.12470

Baalousha, H., 2009. Stochastic water balance model for rainfall recharge quantification in Ruataniwha Basin , New Zealand. Environ. Geol. 58, 85–93. https://doi.org/10.1007/s00254-008-1495-6

Banks, E.W., Hatch, M., Smith, S., Underschultz, J., Lamontagne, S., Suckow, A., Mallants, D., 2019. Multi-tracer and hydrogeophysical investigation of the hydraulic connectivity between coal seam gas formations, shallow groundwater and stream network in a faulted sedimentary basin. J. Hydrol. 578, 124132. https://doi.org/10.1016/j.jhydrol.2019.124132

Barnett, B., Townley, L.R., Post, V.E.A., Evans, R.E., Hunt, R.J., Peeters, L.J.M., Richardson, S., Werner, A.D., Knapton, A., Boronkay, A., 2012. Australian groundwater modelling guidelines, Waterlines Report Series. National Water Commision, Canberra.

Betini, G.S., Avgar, T., Fryxell, J.M., 2017. Why are we not evaluating multiple competing hypotheses in ecology and evolution? R. Soc. Open Sci. 4. https://doi.org/http://dx.doi.org/10.1098/rsos.160756

Beven, K.J., 2018. On hypothesis testing in hydrology: Why falsification of models is still a really good idea. WIREs Water 3, e1278. https://doi.org/10.1002/wat2.1278

Beven, K.J., 2016. Facets of uncertainty: Epistemic uncertainty, non-stationarity, likelihood, hypothesis testing, and communication. Hydrol. Sci. J. 61, 1652–1665. https://doi.org/10.1080/02626667.2015.1031761

Beven, K.J., 2002. Towards a coherent philosophy for environmental modelling. Proc. R. Soc. London A Math. Phys. Eng. Sci. 458, 2465–2484. https://doi.org/10.1098/rspa.2002.0986

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

Beven, K.J., Young, P., 2013. A guide to good practice in modeling semantics for authors and referees. Water Resour. Res. 49, 5092–5098. https://doi.org/10.1002/wrcr.20393

Bird, A., 2018. Thomas Kuhn, in: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University.

Bond, C.E., Gibbs, A.D., Shipton, Z.K., Jones, S., 2007. What do you think this is? “Conceptual uncertainty” in geoscience interpretation. GSA Today 17, 4–10. https://doi.org/10.1130/GSAT01711A.1

Brassington, F.C., Younger, P.L., 2010. A proposed framework for hydrogeological conceptual modelling. Water Environ. 24, 261–273. https://doi.org/10.1111/j.1747-6593.2009.00173.x

Bredehoeft, J.D., 2005. The conceptualization model problem - Surprise. Hydrogeol. J. 13, 37–46. https://doi.org/10.1007/s10040-004-0430-5

Bresciani, E., Cranswick, R.H., Banks, E.W., Batlle-Aguilar, J., Cook, P.G., Batelaan, O., 2018. Using hydraulic head, chloride and electrical conductivity data to distinguish between mountain-front and mountain-block recharge to basin aquifers. Hydrol. Earth Syst. Sci. 22, 1629–1648. https://doi.org/10.5194/hess-22-1629-2018

Brie, A., Pampuri, F., Marsala, A.F., Meazza, O., 1995. Shear sonic interpretation in gas-bearing sands. Proc. - SPE Annu. Tech. Conf. Exhib. Omega, 701–710. https://doi.org/10.2523/30595-ms

Brunetti, C., Linde, N., Vrugt, J.A., 2017. Bayesian model selection in hydrogeophysics: Application to conceptual subsurface models of the South Oyster Bacterial Transport. Adv. Water Resour. 102, 127–141. https://doi.org/10.1016/j.advwatres.2017.02.006

Caers, J., 2018. Bayesianism in Geoscience, in: Sagar, B.S.D., Cheng, Q., Agterberg, F. (Eds.), Handbook of Mathematical Geosciences. Springer, Cham, Stanford University, USA, pp. 527–566. https://doi.org/https://doi.org/10.1007/978-3-319-78999-6_27

Carrera, J., Mousavi, S.F., Usunoff, E.J., Sánchez-Vila, X., Galarza, G., 1993. A discussion on validation of hydrogeological models. Reliab. Eng. Syst. Saf. 42, 201–216. https://doi.org/10.1016/0951-8320(93)90089-H

Carrera, J., Neuman, S.P., 1986. Estimation of Aquifer Parameters Under Transient and Steady State Conditions: 3. Application to Synthetic Field data. Water Resour. Res. 22, 228–242.

Castro, M.C., Goblet, P., 2003. Calibration of regional groundwater flow models: Working toward a better understanding of site-specific systems. Water Resour. Res. 39, 1172. https://doi.org/10.1029/2002WR001653

Chatfield, C., 1995. Model Uncertainty, Data Mining and Statistical Inference. J. R. Stat. Soc. 158, 419–466.

Chitsazan, N., Nadiri, A.A., Tsai, F.T.C., 2015. Prediction and structural uncertainty analyses of artificial neural networks using hierarchical Bayesian model averaging. J. Hydrol. 528, 52–62. https://doi.org/10.1016/j.jhydrol.2015.06.007

Clark, M.P., Kavetski, D., Fenicia, F., 2011. Pursuing the method of multiple working hypotheses for hydrological modeling. Water Resour. Res. 47, 1–16. https://doi.org/10.1029/2010WR009827

Clark, M.P., Slater, A.G., Rupp, D.E., Woods, R.A., Vrugt, J.A., Gupta, H. V, Wagener, T., Hay, L.E., 2008. Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models. Water Resour. Res. 44, 1–14. https://doi.org/10.1029/2007WR006735

Cook, P.G., Bohlke, J.-K., 2000. Determining Timescales for Groundwater Flow and Solute

Transport, in: Cook, P.G., Herczeg, A.L. (Eds.), Environmental Tracers in Subsurface Hydrology. pp. 1–30.

Currell, M.J., Werner, A.D., McGrath, C., Webb, J.A., Berkman, M., 2017. Problems with the application of hydrogeological science to regulation of Australian mining projects: Carmichael Mine and Doongmabulla Springs. J. Hydrol. 548, 674–682. https://doi.org/10.1016/j.jhydrol.2017.03.031

Dassargues, A., 2018. Chapter 2: Hydrologic balance and groundwater, in: Hydrogeology: Groundwater Science and Engineering. CRC Press, Boca Raton, Florida.

Dausman, A.M., Doherty, J., Langevin, C.D., Dixon, J., 2010. Hypothesis testing of buoyant plume migration using a highly parameterized variable-density groundwater model at a site in Florida, USA. Hydrogeol. J. 18, 147–160. https://doi.org/10.1007/s10040-009-0511-6

Davis, P.A., Olague, N.E., Goodrich, M.T., 1991. Approaches for the validation of models used for performance assessment of high-level nuclear waste repositories, SAND90-0575/NUREGC R-5537. Sandia National Laboratories, Albuquerque, NM.

de Graaf, I.E.M., Gleeson, T., (Rens) van Beek, L.P.H., Sutanudjaja, E.H., Bierkens, M.F.P., 2019. Environmental flow limits to global groundwater pumping. Nature 574, 90–94. https://doi.org/10.1038/s41586-019-1594-4

Dickson, N.E.M., Comte, J.-C., Renard, P., Straubhaar, J.A., Mckinley, J.M., Ofterdinger, U., 2015. Integrating aerial geophysical data in multiple-point statistics simulations to assist groundwater flow models. Hydrogeol. J. 23, 883–900. https://doi.org/10.1007/s10040-015-1258-x

Diks, C.G.H., Vrugt, J.A., 2010. Comparison of point forecast accuracy of model averaging methods in hydrologic applications. Stoch. Environ. Res. Risk Assess. 24, 809–820. https://doi.org/10.1007/s00477-010-0378-z

Dixon, T.H., Amelung, F., Ferretti, A., Novali, F., Rocca, F., Dokka, R., Sella, G., Kim, S.-W., Wdowinski, S., Whitman, D., 2006. Subsidence and flooding in New Orleans. Nature 441, 587–588.

Doble, R.C., Crosbie, R.S., 2017. Review: Current and emerging methods for catchment-scale modelling of recharge and evapotranspiration from shallow groundwater. Hydrogeol. J. 25, 3–23. https://doi.org/10.1007/s10040-016-1470-3

Doherty, J., Welter, D., 2010. A short exploration of structural noise. Water Resour. Res. 46, 1–14. https://doi.org/10.1029/2009WR008377

Döll, P., Fiedler, K., 2008. Global-scale modeling of groundwater recharge. Hydrol. Earth Syst. Sci. 12, 863–885. https://doi.org/10.5194/hess-12-863-2008

Dvorkin, J., Prasad, M., Sakai, A., Lavoie, D., 1999. Elasticity of marine sediments: Rock physics modeling. Geophys. Res. Lett. 26, 1781–1784. https://doi.org/10.1029/1999GL900332

Easey, D., Brocklehurst, P., Carnavas, M., 2016. Agricultural Land Suitability Series, Report 2. Soil and Land Suitability Assessment for Irrigated Agriculture in the Wildman River area, Northern Territory. Darwin, NT, Australia.

Eckhardt, K., 2008. A comparison of baseflow indices , which were calculated with seven different baseflow separation methods. J. Hydrol. 352, 168–173. https://doi.org/10.1016/j.jhydrol.2008.01.005

Eckhardt, K., 2005. How to construct recursive digital filters for baseflow separation. Hydrol. Process. 19, 507–515. https://doi.org/10.1002/hyp.5675

Elshall, A.S., Tsai, F.T.C., 2014. Constructive epistemic modeling of groundwater flow with geological structure and boundary condition uncertainty under the Bayesian paradigm. J. Hydrol. 517, 105–119. https://doi.org/10.1016/j.jhydrol.2014.05.027

Elshall, A.S., Tsai, F.T.C., Hanor, J.S., 2013. Indicator geostatistics for reconstructing Baton Rouge aquifer-fault hydrostratigraphy, Louisiana, USA. Hydrogeol. J. 21, 1731–1747. https://doi.org/10.1007/s10040-013-1037-5

Enemark, T., Peeters, L.J.M., Mallants, D., Batelaan, O., 2019a. Hydrogeological conceptual model building and testing: A review. J. Hydrol. 569. https://doi.org/10.1016/j.jhydrol.2018.12.007

Enemark, T., Peeters, L.J.M., Mallants, D., Batelaan, O., Valentine, A.P., Sambridge, M., 2019b. Hydrogeological Bayesian Hypothesis Testing through Trans-Dimensional Sampling of a Stochastic. Water 11. https://doi.org/10.3390/w11071463

Engelhardt, I., De Aguinaga, J.G., Mikat, H., Schüth, C., Liedl, R., 2014. Complexity vs. Simplicity: Groundwater Model Ranking Using Information Criteria. Groundwater 52, 573–583. https://doi.org/10.1111/gwat.12080

Environment Agency, 2002. Groundwater resources modelling: guidance notes and template project brief, Environment Agency R&D Guidance Notes W213. Environment Agency, Bristol.

Ferré, T.P.A., 2017. Revisiting the Relationship Between Data, Models, and Decision-Making. Groundwater 55, 604–614. https://doi.org/10.1111/gwat.12574

Feyen, L., Caers, J., 2006. Quantifying geological uncertainty for flow and transport modeling in multi-modal heterogeneous formations. Adv. Water Resour. 29, 912–929. https://doi.org/10.1016/j.advwatres.2005.08.002

Feyen, L., Gorelick, S.M., 2005. Framework to evaluate the worth of hydraulic conductivity data for optimal groundwater resources management in ecologically sensitive areas. Water Resour. Res. 41, 1–13. https://doi.org/10.1029/2003WR002901

Fisher, R.A., 1935. The factorial design of experimentation, in: The Design of Experiments. Oliver and Boyd, London, United Kingdom, pp. 96–113.

Flinchum, B.A., Holbrook, W.S., Grana, D., Parsekian, A.D., Carr, B.J., Hayes, J.L., Jiao, J., 2018. Estimating the water holding capacity of the critical zone using near-surface geophysics. Hydrol. Process. 32, 3308–3326. https://doi.org/10.1002/hyp.13260

Flint, A.L., Flint, L.E., Hevesi, J.A., D'Agnese, F., Faunt, C., 2000. Estimation of regional recharge and travel time through the unsaturated zone in arid climates. Geophys. Monogr. Ser. 122, 115–128. https://doi.org/10.1029/GM122p0115

Foglia, L., Mehl, S.W., Hill, M.C., Burlando, P., 2013. Evaluating model structure adequacy: The case of the Maggia Valley groundwater system, southern Switzerland. Water Resour. Res. 49, 260–282. https://doi.org/10.1029/2011WR011779

Foglia, L., Mehl, S.W., Hill, M.C., Perona, P., Burlando, P., 2007. Testing alternative ground water models using cross-validation and other methods. Ground Water 45, 627–641. https://doi.org/10.1111/j.1745-6584.2007.00341.x

Francés, A.P., Lubczynski, M.W., Roy, J., Santos, F.A.M., Mahmoudzadeh Ardekani, M.R., 2014. Hydrogeophysics and remote sensing for the design of hydrogeological conceptual models in hard rocks - Sardón catchment (Spain). J. Appl. Geophys. 110, 63–81. https://doi.org/10.1016/j.jappgeo.2014.08.015

Freedman, V.L., Truex, M.J., Rockhold, M.L., Bacon, D.H., Freshley, M.D., Wellman, D.M., 2017. Elements of complexity in subsurface modeling, exemplified with three case studies. Hydrogeol. J. 25, 1853–1870. https://doi.org/10.1007/s10040-017-1564-6

Gedeon, M., Mallants, D., Rogiers, B., 2013. Building a staircase of confidence in groundwater modeling: a summary of ten years data collection and model development, in: Modflow and More Conference: Translating Science into Practice. Golden, CO.

Gelman, A., Shalizi, C.R., 2013. Philosophy and the practice of Bayesian statistics 8–38. https://doi.org/10.1111/j.2044-8317.2011.02037.x

Giordano, M., 2009. Global Groundwater? Issues and Solutions. Annu. Rev. Environ. Resour. 34, 153–178. https://doi.org/10.1146/annurev.environ.030308.100251

Gondwe, B.R.N., Lerer, S., Stisen, S., Marín, L., Rebolledo-Vieyra, M., Merediz-Alonso, G., Bauer-Gottwein, P., 2010. Hydrogeology of the south-eastern Yucatan Peninsula: New insights from water level measurements, geochemistry, geophysics and remote sensing. J. Hydrol. 389, 1–17. https://doi.org/10.1016/j.jhydrol.2010.04.044

Graham, B., 1985. Surface water resources in the northeastern corner of Wildman River Station (Final report for Water Resources Division pOliphant, T. E. (2006). A Guide to NumPy. Tregol Publishing.roject number 2026). Darwin, Northern Territory, Australia.

Green, P.J., 2003. Trans-dimensional Markov chain Monte Carlo, in: Green, P.J., Hjort, N.L., Richardson, S. (Eds.), Highly Structured Stochastic Systems. Oxford Statistical Science Series, pp. 179–198.

Green, P.J., 1995. Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. Biometrika 82, 711–732. https://doi.org/10.2307/2337340

Guillaume, J.H.A., Hunt, R.J., Comunian, A., Blakers, R.S., Fu, B., 2016. Methods for Exploring Uncertainty in Groundwater Management Predictions, in: Jakeman, A.J., Barreteau, O., Hunt, R.J., Rinaudo, J., Ross, A. (Eds.), Integrated Groundwater

Management. Springer, Cham, pp. 602–614. https://doi.org/https://doi.org/10.1007/978-3-319-23576-9_28

Gupta, H. V, Clark, M.P., Vrugt, J.A., Abramowitz, G., Ye, M., 2012. Towards a comprehensive assessment of model structural adequacy. Water Resour. Res. 48, 1–16. https://doi.org/10.1029/2011WR011044

Haitjema, H.M., 2006. The Role of Hand Calculations in Ground Water Flow Modeling. Groundwater 44, 786–791. https://doi.org/10.1111/j.1745-6584.2006.00189.x

Haitjema, H.M., 1995. Introduction, in: Analytic Element Modeling of Groundwater Flow. Academic Press, pp. 1–4.

Harrar, W.G., Sonnenborg, T.O., Henriksen, H.J., 2003. Capture zone, travel time, and solute-transport predictions using inverse modeling and different geological models. Hydrogeol. J. 11, 536–548. https://doi.org/10.1007/s10040-003-0276-2

Hashin, Z., Shtrikman, S., 1963. A variational approach to the theory of the elastic behaviour of multiphase materials. J. Mech. Phys. Solids 11, 127–140.

Hassan, A.E., 2004. Validation of Numerical Ground Water Models Used to Guide Decision Making. Groundwater 42, 277–290.

Hassan, A.E., 2003. A Validation Process for the Groundwater Flow and Transport Model of the Faultless Nuclear Test at Central Nevada Test Area, Division of Hydrologic Sciences Publication, No. 45197. Las Vegas, Nevada, USA. https://doi.org/10.2172/812127

Hastings, W.K., 1970. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika 57, 97–109.

He, X., Sonnenborg, T.O., Jørgensen, F., Jensen, K.H., 2014. The effect of training image and secondary data integration with multiple-point geostatistics in groundwater modelling.

Hydrol. Earth Syst. Sci. 18, 2943–2954. https://doi.org/10.5194/hess-18-2943-2014

Helgerud, M.B., 2001. Wave speeds in gas hydrate and sediments containing gas hydrate: a laboratory and modelling study. Dep. Geophys. 249.

Helgerud, M.B., Dvorkin, J., Nur, A., Sakai, A., Collett, T., 1999. Elastic-wave velocity in marine sediments with gas hydrates: Effective medium modeling. Geophys. Res. Lett. 26, 2021–2024. https://doi.org/10.1029/1999GL900421

Hermans, T., Nguyen, F., Caers, J., 2015. Uncertainty in training image-based inversion of hydraulic head data constrained to ERT data: Workflow and case study. Water Resour. Res. 51, 5332–5352. https://doi.org/10.1002/ 2014WR016460

Hills, R.G., Wierenga, P.J., 1994. INTRAVAL Phase II Model Testing at the Las Cruces Trench Site. NUREG/CR-6063.

Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian Model Averaging: A Tutorial. Stat. Sci. 14, 382–417. https://doi.org/10.2307/2676803

Höge, M., Guthke, A., Nowak, W., 2019. The hydrologist’s guide to Bayesian model selection, averaging and combination. J. Hydrol. 572, 96–107. https://doi.org/10.1016/j.jhydrol.2019.01.072

Höge, M., Wöhling, T., Nowak, W., 2018. A Primer for Model Selection: The Decisive Role of Model Complexity. Water Resour. Res. 1688–1715. https://doi.org/10.1002/2017WR021902

Højberg, A.L., Refsgaard, J.C., 2005. Model uncertainty - parameter uncertainty versus conceptual models. Water Sci. Technol. 52, 177–186.

Holbrook, W.S., Riebe, C.S., Elwaseif, M., Hayes, J.L., Basler-Reeder, K., Harry, D.L., Malazian, A., Dosseto, A., Hartsough, P.C., Hopmans, J.W., 2014. Geophysical constraints on deep weathering and water storage potential in the Southern Sierra Critical Zone Observatory. Earth Surf. Process. Landforms 39, 366–380. https://doi.org/10.1002/esp.3502

Hunt, R.J., Welter, D.E., 2010. Taking Account of “Unknown Unknowns.” Ground Water 48, 477–477. https://doi.org/10.1111/j.1745-6584.2010.00681.x

Hunt, R.J., Zheng, C., 2012. The Current State of Modeling. Ground Water 50, 330–333. https://doi.org/10.1111/j.1745-6584.2012.00936.x

Hunter, J.D., 2007. Matplotlib: A 2D Graphics Environment. Comput. Sci. Eng. 9, 90–95. https://doi.org/citeulike-article-id:2878517\rdoi: 10.1109/mcse.2007.55

Hurvich, C.M., Tsai, C.-L., 1989. Regression and Time Series Model Selection in Small Samples, Biometrika. https://doi.org/10.1093/biomet/76.2.297

Izady, A., Davary, K., Alizadeh, A., Ziaei, A.N., Alipoor, A., Joodavi, A., Brusseau, M.L., 2014. A framework toward developing a groundwater conceptual model. Arab. Jounal Geosci. 7, 3611–3631. https://doi.org/10.1007/s12517-013-0971-9

Jakeman, A.J., Letcher, R.A., Norton, J.P., 2006. Ten iterative steps in development and evaluation of environmental models. Environ. Model. Softw. 21, 602–614. https://doi.org/10.1016/j.envsoft.2006.01.004

Jardani, A., Revil, A., Santos, F., Fauchard, C., Dupont, J.P., 2007. Detection of preferential infiltration pathways in sinkholes using joint inversion of self-potential and EM-34 conductivity data. Geophys. Prospect. 55, 749–760. https://doi.org/10.1111/j.1365-2478.2007.00638.x

Jeffreys, H., 1939. Theory of Probability, third edit. ed. Oxford University Press.

Jiménez, S., Mariethoz, G., Brauchler, R., Bayer, P., 2016. Smart pilot points using reversible-jump Markov-chain Monte Carlo. Water Resour. Res. 52, 3966–3983. https://doi.org/10.1002/2015WR017922

Johnson, G.S., Frederick, D.B., Cosgrove, D.M., 2002. Evaluation of a pumping test of the Snake River Plain aquifer using axial-flow numerical modeling. Hydrogeol. J. 10, 428–437. https://doi.org/10.1007/s10040-002-0201-0

Kass, R.E., Raftery, A.E., 1995. Bayes Factors. J. ofthe Am. Stat. Assoc. 90, 773–795.

Kerr, N.L., 1998. HARKing: Hypothesizing After the Results are Known. Personal. Soc. Psychol. Rev. 2, 196–217. https://doi.org/10.1207/s15327957pspr0203

Kikuchi, C.P., Ferré, T.P.A., Vrugt, J.A., 2015. On the optimal design of experiments for conceptual and predictive discrimination of hydrologic system models. Water Resour. Res. 4454–4481. https://doi.org/10.1002/2014WR016795

Knopman, D.S., Voss, C.I., 1989. Multiobjective Sampling Design for Parameter Estimation and Model Discrimination in Groundwater Solute Transport. Water Resour. Res. 25, 2245–2258.

Knopman, D.S., Voss, C.I., 1988. Discrimination among one-dimensional models of solute transport in porous media: Implications for sampling design. Water Resour. Res. 24, 1859–1876. https://doi.org/10.1029/WR024i011p01859

Knopman, D.S., Voss, C.I., Garabedian, S.P., 1991. Sampling Design for Groundwater Solute Transport - Tests of Methods and Analysis of Cape-Cod Tracer Test Data. Water Resour. Res. 27, 925–949.

Kohavi, R., 1995. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, in: International Joint Conference on Articial Intelligence (IJCAI). Montreal, Canada, pp. 1137–1145. https://doi.org/10.1067/mod.2000.109031

Konikow, L.F., Bredehoeft, J.D., 1992. Ground-water models cannot be validated. Adv. Water Resour. 15, 75–83. https://doi.org/10.1016/0309-1708(92)90033-X

Konikow, L.F., Kendy, E., 2005. Groundwater depletion: A global problem. Hydrogeol. J. 13, 317–320. https://doi.org/10.1007/s10040-004-0411-8

Krabbenhoft, D.P., Anderson, M.P., 1986. Use of a numerical ground-water flow model for hypothesis testing. Ground Water 24, 49–55.

Krueger, T., Page, T., Hubacek, K., Smith, L., Hiscock, K., 2012. The role of expert opinion in environmental modelling. Environ. Model. Softw. 36, 4–18. https://doi.org/10.1016/j.envsoft.2012.01.011

Kruse, S., Grasmueck, M., Weiss, M., Viggiano, D., 2006. Sinkhole structure imaging in covered Karst terrain. Geophys. Res. Lett. 33, 1–6. https://doi.org/10.1029/2006GL026975

La Vigna, F., Demiray, Z., Mazza, R., 2014. Exploring the use of alternative groundwater models to understand the hydrogeological flow processes in an alluvial context (Tiber River, Rome, Italy). Envrionment Earth Sci. 71, 1115–1121. https://doi.org/10.1007/s12665-013-2515-8

Lee, J., Sung, W., Choi, J.H., 2015. Metamodel for efficient estimation of capacity-fade uncertainty in Li-Ion batteries for electric vehicles. Energies 8, 5538–5554. https://doi.org/10.3390/en8065538

Lee, R.R., Ketelle, R.H., Bownds, J.M., Rizk, T.A., 1992. Aquifer Analysis and Modeling in a Fractured Heterogeneous Medium. Ground Water 30, 589–597.

Leterme, B., Mallants, D., Jacques, D., 2012. Sensitivity of groundwater recharge using climatic analogues and HYDRUS-1D. Hydrol. Earth Syst. Sci. 16, 2485–2497.

https://doi.org/10.5194/hess-16-2485-2012

Li, X., Tsai, F.T.C., 2009. Bayesian model averaging for groundwater head prediction and uncertainty analysis using multimodel and multimethod. Water Resour. Res. 45, 1–14. https://doi.org/10.1029/2008WR007488

Linde, N., 2014. Falsification and corroboration of conceptual hydrological models using geophysical data. Wiley Interdiscip. Rev. Water 1, 151–171. https://doi.org/10.1002/wat2.1011

Linde, N., Lochbühler, T., Dogan, M., Van Dam, R.L., 2015a. Tomogram-based comparison of geostatistical models: Application to the Macrodispersion Experiment (MADE) site. J. Hydrol. 531, 543–556. https://doi.org/10.1016/j.jhydrol.2015.10.073

Linde, N., Renard, P., Mukerji, T., Caers, J., 2015b. Geological realism in hydrogeological and geophysical inverse modeling: A review. Adv. Water Resour. 86, 86–101. https://doi.org/10.1016/j.advwatres.2015.09.019

Lloyd, D., 1999. Seasonal Changes in Two Tropical Freshwater Lagoons of Northern Australia. Nat. Resour. Div.

Lukjan, A., Swasdi, S., Chalermyanont, T., 2016. Importance of Alternative Conceptual Model for Sustainable Groundwater Management of the Hat Yai Basin, Thailand. Procedia Eng. 154, 308–316. https://doi.org/10.1016/j.proeng.2016.07.480

Lyne, V., Hollick, M., 1979. Stochastic Time-Variable Rainfall-Runoff Modeling, in: Institute of Engineers Australia National Conference. pp. 89–93.

Malinverno, A., Leaney, W.., 2000. A Monte Carlo method to quantify uncertainty in the inversion of zero-offset vsp data, in: 70th SEG Annual Meeting Expanded Abstracts. Tulsa, Oklahoma, pp. 2392–2396.

Markou, M., Singh, S., 2003. Novelty detection: A review - Part 1: Statistical approaches. Signal Processing 83, 2481–2497. https://doi.org/10.1016/j.sigpro.2003.07.018

Marshall, L., 2017. Creativity, Uncertainty, and Automated Model Building. Groundwater. https://doi.org/10.1111/gwat.12552

Martinez, G.F., Gupta, H. V, 2011. Hydrologic consistency as a basis for assessing complexity of monthly water balance models for the continental United States. Water Resour. Res. 47, 1–18. https://doi.org/10.1029/2011WR011229

Mavko, G., Mukerji, T., Dvorkin, J., 2009. The rock physics handbook tools for seismic analysis of porous media. New York: Cambridge University Press, Cambridge, UK.

McFeeters, S.K., 1996. The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. Int. J. Remote Sens. 17, 1425–1432. https://doi.org/https://doi.org/10.1080/01431169608948714

Mechal, A., Birk, S., Winkler, G., Wagner, T., Mogessie, A., 2016. Characterizing regional groundwater flow in the Ethiopian Rift : A multi- model approach applied to Gidabo River Basin. Austrian J. Earth Sci. 109. https://doi.org/10.17738/ajes.2016.0005

Metropolis, N., Ulam, S., 1949. The Monet Carlo Method. J. Am. Stat. Assoc. 44, 335–341.

Meyer, P., Gee, G., 1999. Information on hydrologic conceptual models, parameters, uncertainty analysis, and data sources for dose assessments at decommissioning sites, NUREG/CR-6656. Washington, D.C.

Meyer, P.D., Ye, M., Neuman, S.P., Cantrell, K.J., 2003. Combined Estimation of Hydrogeologic Conceptual Model and Parameter Uncertainty. NUREG/CR-6843 Report. Washington, D.C.

Meyer, P.D., Ye, M., Rockhold, M.L., Neuman, S.P., Cantrell, K.J., 2007. Combined

Estimation of Hydrogeologic Conceptual Model , Parameter , and Scenario Uncertainty with Application to Uranium Transport at the Hanford Site 300 Area. US Nucl. Regul. Commision NUREG/CR-6.

Middlemis, H., Walker, G., Peeters, L.J., Hayes, P., Moore, C., 2019. Groundwater modelling uncertainty – implications for decision making. Summary report of the national groundwater modelling uncertainty workshop, 10 July 2017, Sydney, Australia. https://doi.org/10.25957/5ca5641defe56

Mindlin, R.D., 1949. Compliance of elastic bodies in contact. J. Appl. Mech. 16, 259–268.

Mineralogy Database [WWW Document], n.d. URL http://www.webmineral.com/ (accessed 9.27.19).

Mondal, A., Efendiev, Y., Mallick, B., Datta-Gupta, A., 2010. Bayesian uncertainty quantification for flows in heterogeneous porous media using reversible jump Markov chain Monte Carlo methods. Adv. Water Resour. 33, 241–256. https://doi.org/10.1016/j.advwatres.2009.10.010

Moore, C., Doherty, J., 2005. Role of the calibration process in reducing model predictive error. Water Resour. Res. 41, 1–14. https://doi.org/10.1029/2004WR003501

Moser, T.J., 1991. Shortest path calculation of seismic rays. Geophysics 56, 59–67.

Mueller, N., Lewis, A., Roberts, D., Ring, S., Melrose, R., Sixsmith, J., Lymburner, L., McIntyre, A., Tan, P., Curnow, S., Ip, A., 2016. Water observations from space: Mapping surface water from 25years of Landsat imagery across Australia. Remote Sens. Environ. 174, 341–352. https://doi.org/10.1016/j.rse.2015.11.003

Mustafa, S.M.T., Nossent, J., Ghysels, G., Huysmans, M., 2020. Integrated Bayesian Multi-model approach to quantify input, parameter and conceptual model structure uncertainty in groundwater modeling. Environ. Model. Softw. 126. https://doi.org/10.1016/j.envsoft.2020.104654

Nearing, G.S., Gupta, H. V, 2018. Ensembles vs. information theory: supporting science under uncertainty. Front. Earth Sci. 1–8. https://doi.org/10.1007/s11707-018-0709-9

Nearing, G.S., Tian, Y., Gupta, H. V, Clark, M.P., Harrison, K.W., Weijs, S. V, 2016. A philosophical basis for hydrological uncertainty. Hydrol. Sci. J. 61, 1666–1678. https://doi.org/10.1080/02626667.2016.1183009

Neto, D.C., Chang, H.K., van Genuchten, M.T., 2016. A Mathematical View of Water Table Fluctuations in a Shallow Aquifer in Brazil. Groundwater 54, 82–91. https://doi.org/10.1111/gwat.12329

Nettasana, T., 2012. Conceptual Model Uncertainty in the Management of the Chi River Basin, Thailand. University of Waterloo, PhD Thesis.

Nettasana, T., Craig, J., Tolson, B., 2012. Conceptual and numerical models for sustainable groundwater management in the Thaphra area, Chi River Basin, Thailand. Hydrogeol. J. 20, 1355–1374. https://doi.org/10.1007/s10040-012-0887-6

Neuman, S.P., 2003. Maximum likelihood Bayesian averaging of uncertain model predictions. Stoch. Environ. Res. Risk Assess. 17, 291–305. https://doi.org/10.1007/s00477-003-0151-7

Neuman, S.P., Wierenga, P.J., 2003. A Comprehensive Strategy of Hydrogeologic Modeling and Uncertainty Analysis for Nuclear Facilities and Sites. NUREG/CR-6805 311.

Nishikawa, T., 1997. Testing alternative conceptual models of seawater intrusion in a coastal aquifer using computer simulation, southern California, USA. Hydrogeol. J. https://doi.org/10.1007/s100400050116

Nordqvist, R., Voss, C.I., 1996. A simulation-based approach for designing effective field-sampling programs to evaluate contamination risk of groundwater supplies. Hydrogeol. J. 4, 23–39.

Nur, A., Mavko, G., Dvorkin, J., Gal, D., 1998. Critical porosity: The key to relating physical properties to porosity in rocks. Lead. Edge 357–362. https://doi.org/10.1190/1.1887540

Okasha, S., 2002. Philosophy of Science: A Very Short Introduction. Oxford University Press, Oxford.

Oliphant, T.E., 2006. A Guide to NumPy. Tregol Publishing.

Oreskes, N., Shrader-frechette, K., Belitz, K., 1994. Verification , Validation and Confirmation of Numerical Models in the Earth Sciences. Science (80-.). 263, 641–646.

Othman, A., Sultan, M., Becker, R., Alsefry, S., Alharbi, T., Gebremichael, E., Alharbi, H., Abdelmohsen, K., 2018. Use of Geophysical and Remote Sensing Data for Assessment of Aquifer Depletion and Related Land Deformation. Surv. Geophys. 39, 543–566. https://doi.org/10.1007/s10712-017-9458-7

Otsu, N., 1979. A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. 9, 62–66. https://doi.org/doi:10.1117/1.1631315

Pappenberger, F., Beven, K.J., 2006. Ignorance is bliss: Or seven reasons not to use uncertainty analysis. Water Resour. Res. 42, 1–8. https://doi.org/10.1029/2005WR004820

Park, H., Scheidt, C., Fenwick, D., Boucher, A., Caers, J., 2013. History matching and uncertainty quantification of facies models with multiple geological interpretations. Comput. Geosci. 17, 609–621. https://doi.org/10.1007/s10596-013-9343-5

Passadore, G., Monego, M., Altissimo, L., Sottani, A., Putti, M., 2011. Alternative conceptual models and the robustness of groundwater management scenarios in the multi-aquifer system of the Central Veneto Basin , Italy. https://doi.org/10.1007/s10040-011-0818-y

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Peeters, L.J.M., 2017. Assumption Hunting in Groundwater Modeling: Find Assumptions Before They Find You. Groundwater. https://doi.org/10.1111/gwat.12565

Peeters, L.J.M., Crosbie, R.S., Doble, R.C., Van Dijk, A.I.J.M., 2013. Conceptual evaluation of continental land-surface model behaviour. Environ. Model. Softw. 43, 49–59. https://doi.org/10.1016/j.envsoft.2013.01.007

Perko, J., Seetharam, S.C., Mallants, D., Vermariën, E., Wilmot, R., 2009. Long-term evolution of the near surface disposal facility at Dessel. Project near surface disposal of category A waste at Dessel.

Pfister, L., Kirchner, J.W., 2017. Debates - Hypothesis testing in hydrology: Theory and practice. Water Resour. Res. 53, 1792–1798. https://doi.org/10.1002/2016WR020116.Received

Pham, H. V., Tsai, F.T.C., 2016. Optimal observation network design for conceptual model discrimination and uncertainty reduction. Water Resour. Res. 52, 1245–1264. https://doi.org/10.1002/2015WR017474

Pham, H. V., Tsai, F.T.C., 2015. Bayesian experimental design for identification of model propositions and conceptual model uncertainty reduction. Adv. Water Resour. 83, 148–159. https://doi.org/10.1016/j.advwatres.2015.05.024

Pielke, R.A., 2001. Room for doubt. Nature 410, 151. https://doi.org/10.1038/35065759

Pirot, G., Huber, E., Irving, J., Linde, N., 2019. Reduction of conceptual model uncertainty using ground-penetrating radar profiles: Field-demonstration for a braided-river aquifer. J. Hydrol. 571, 254–264. https://doi.org/10.1016/j.jhydrol.2019.01.047

Pirot, G., Renard, P., Huber, E., Straubhaar, J., Huggenberger, P., 2015. Influence of conceptual model uncertainty on contaminant transport forecasting in braided river aquifers. J. Hydrol. 531, 124–141. https://doi.org/10.1016/j.jhydrol.2015.07.036

Planet Team, 2017. Planet Application Program Interface: In Space for Life on Earth. San Francisco, CA.

Poeter, E., Anderson, D., 2005. Multimodel ranking and inference in ground water modeling. Ground Water 43, 597–605. https://doi.org/10.1111/j.1745-6584.2005.0061.x

Post, V.E.A., 2005. Fresh and saline groundwater interaction in coastal aquifers: Is our technology ready for the problems ahead? Hydrogeol. J. 13, 120–123. https://doi.org/10.1007/s10040-004-0417-2

Powell, S.L., Cohen, W.B., Yang, Z., Pierce, J.D., Alberti, M., 2008. Quantification of impervious surface in the Snohomish Water Resources Inventory Area of Western Washington from 1972-2006. Remote Sens. Environ. 112, 1895–1908. https://doi.org/10.1016/j.rse.2007.09.010

Raftery, A.E., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005. Using Bayesian Model Averaging to Calibrate Forecast Ensembles. Mon. Weather Rev. 133, 1155–1174. https://doi.org/10.1175/MWR2906.1

Rajabi, M.M., Ataie-Ashtiani, B., Simmons, C.T., 2018. Model-data interaction in groundwater studies: Review of methods, applications and future directions. J. Hydrol.

567, 457–477. https://doi.org/10.1016/j.jhydrol.2018.09.053

Ray, A., Key, K., 2012. Bayesian inversion of marine CSEM data with a trans-dimensional self parametrizing algorithm. Geophys. J. Int. 191, 1135–1151. https://doi.org/10.1111/j.1365-246X.2012.05677.x

Rayner, D., 2005. Australian synthetic daily Class A pan evaporation. Report 7.

Reeves, D.M., Pohlmann, K.F., Pohll, G.M., Ye, M., Chapman, J.B., 2010. Incorporation of conceptual and parametric uncertainty into radionuclide flux estimates from a fractured granite rock mass. Stoch. Environ. Res. Risk Assess. 24, 899–915. https://doi.org/10.1007/s00477-010-0385-0

Refsgaard, J.C., Christensen, S., Sonnenborg, T.O., Seifert, D., Højberg, A.L., Troldborg, L., 2012. Review of strategies for handling geological uncertainty in groundwater flow and transport modeling. Adv. Water Resour. 36, 36–50. https://doi.org/10.1016/j.advwatres.2011.04.006

Refsgaard, J.C., van der Sluijs, J.P., Brown, J., van der Keur, P., 2006. A framework for dealing with uncertainty due to model structure error. Adv. Water Resour. 29, 1586–1597. https://doi.org/10.1016/j.advwatres.2005.11.013

Remson, I., Gorelick, S.M., Fliegner, J.F., 1980. Computer Models in Ground-Water Exploration. Ground Water 18.

Rogiers, B., Vienken, T., Gedeon, M., Batelaan, O., Mallants, D., Huysmans, M., Dassargues, A., 2014. Multi-scale aquifer characterization and groundwater flow model parameterization using direct push technologies. Environ. Earth Sci. 72, 1303–1324. https://doi.org/10.1007/s12665-014-3416-1

Rojas, R.M., Batelaan, O., Feyen, L., Dassargues, A., 2010a. Assessment of conceptual model uncertainty for the regional aquifer Pampa del Tamarugal – North Chile. Hydrol. Earth Syst. Sci. Discuss. 6, 5881–5935. https://doi.org/10.5194/hessd-6-5881-2009

Rojas, R.M., Feyen, L., Batelaan, O., Dassargues, A., 2010b. On the value of conditioning data to reduce conceptual model uncertainty in groundwater modeling. Water Resour. Res. 46. https://doi.org/10.1029/2009WR008822

Rojas, R.M., Feyen, L., Dassargues, A., 2009. Sensitivity analysis of prior model probabilities and the value of prior knowledge in the assessment of conceptual model uncertainty in groundwater modelling. Hydrol. Process. 23, 1131–1146. https://doi.org/10.1002/hyp

Rojas, R.M., Feyen, L., Dassargues, A., 2008. Conceptual model uncertainty in groundwater modeling: Combining generalized likelihood uncertainty estimation and Bayesian model averaging. Water Resour. Res. 44. https://doi.org/10.1029/2008WR006908

Rojas, R.M., Kahunde, S., Peeters, L.J.M., Batelaan, O., Feyen, L., Dassargues, A., 2010c. Application of a multimodel approach to account for conceptual model and scenario uncertainties in groundwater modelling. J. Hydrol. 394, 416–435. https://doi.org/10.1016/j.jhydrol.2010.09.016

Rosin, P.L., 2001. Unimodal thresholding. Pattern Recognit. 34, 2083–2096. https://doi.org/10.1016/S0031-3203(00)00136-9

Rücker, C., Günther, T., Wagner, F.M., 2017. pyGIMLi: An open-source library for modelling and inversion in geophysics. Comput. Geosci. 109, 106–123. https://doi.org/10.1016/j.cageo.2017.07.011

Samani, S., Moghaddam, A.A., Ye, M., 2017. Investigating the effect of complexity on groundwater flow modeling uncertainty. Stoch. Environ. Res. Risk Assess. 643–659. https://doi.org/10.1007/s00477-017-1436-6

Sambridge, M., Bodin, T., Gallagher, K., Tkalcic, H., 2013. Transdimensional inference in the geosciences. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. 371. https://doi.org/10.1098/rsta.2011.0547

Sambridge, M., Gallagher, K., Jackson, A., Rickwood, P., 2006. Trans-dimensional inverse problems, model comparison and the evidence. Geophys. J. Int. 167, 528–542. https://doi.org/10.1111/j.1365-246X.2006.03155.x

Samper, F.J., Neuman, S.P., 1989. Estimation of Spatial Covariance Structures by Adjoint. Water Resour. Res. 25, 373–384.

Sanford, W.E., Buapeng, S., 1996. Assesment of a Groundwater Flow Model of the Bangkok Basin, Thailand using Carbon-14-based Ages and Paleohydrology. Hydrogeol. J. 4.

Scanlon, B.R., Healy, R.W., Cook, P.G., 2002. Choosing appropriate techniques for quantifying groundwater recharge. Hydrogeol. J. 10, 18–39. https://doi.org/10.1007/s10040-0010176-2

Scheidt, C., Jeong, C., Mukerji, T., Caers, J., 2015. Probabilistic falsification of prior geologic uncertainty with seismic amplitude data: Application to a turbidite reservoir case. Geophysics 80, M89–M12. https://doi.org/10.1190/geo2015-0084.1

Schickore, J., 2018. Scientific Discovery, in: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University.

Schöniger, A., Illman, W.A., Wöhling, T., Nowak, W., 2015a. Finding the right balance between groundwater model complexity and experimental effort via Bayesian model selection. J. Hydrol. 531, 96–110. https://doi.org/10.1016/j.jhydrol.2015.07.047

Schöniger, A., Wöhling, T., Nowak, W., 2015b. A statistical concept to assess the uncertainty in Bayesian model weights and its impact on model ranking. Water Resour. Res. 51, 7524–7546. https://doi.org/10.1002/2015WR016918

Schöniger, A., Wöhling, T., Samaniego, L., Nowak, W., 2014. Model selection on solid ground: Rigorous comparison of nine ways to evaluate Bayesian model evidence. Water Resour. Res. 50, 9484–9513. https://doi.org/10.1002/2014WR016062

Schult, J., Welch, M., 2006. The water quality of fifteen lagoons in the Darwin Region. Report13/2006D. Darwin, Northern Territory, Australia.

Schwartz, F.W., Liu, G., Aggarwal, P., Schwartz, C.M., 2017. Naïve Simplicity: The Overlooked Piece of the Complexity-Simplicity Paradigm. Groundwater 1–9. https://doi.org/10.1111/gwat.12570

Schwarz, G., 1978. Estimating the Dimension of a Model. Ann. Stat. 6, 461–464. https://doi.org/10.1214/aos/1176344136

Sebok, E., Refsgaard, J.C., Warmink, J.J., Stisen, S., Jensen, K.H., 2016. Using expert elicitation to quantify catchment water balances and their uncertainties. Water Resour. Res. 52, 5111–5131. https://doi.org/10.1002/2015WR018461.Received

Seifert, D., Sonnenborg, T.O., Refsgaard, J.C., Højberg, A.L., Troldborg, L., 2012. Assessment of hydrological model predictive ability given multiple conceptual geological models. Water Resour. Res. 48, 1–16. https://doi.org/10.1029/2011WR011149

Seifert, D., Sonnenborg, T.O., Scharling, P.B., Hinsby, K., 2008. Use of alternative conceptual models to assess the impact of a buried valley on groundwater vulnerability. Hydrogeol. J. 16, 659–674. https://doi.org/10.1007/s10040-007-0252-3

Selroos, J.-O., Walker, D.D., Ström, A., Gylling, B., Follin, S., 2002. Comparison of alternative modelling approaches for groundwater flow in fractured rock. J. Hydrol. 257, 174–188. https://doi.org/10.1016/S0022-1694(01)00551-0

Sidiropoulos, E., Tolikas, P., 2004. Well locations and constraint handling in groundwater pumping cost minimization via genetic algorithms. Water, Air, Soil Pollut. Focus 4, 227–239. https://doi.org/10.1023/B:WAFO.0000044801.47725.b6

Singh, A., Mishra, S., Ruskauff, G., 2010. Model averaging techniques for quantifying conceptual model uncertainty. Ground Water 48, 701–715. https://doi.org/10.1111/j.1745-6584.2009.00642.x

Somogyvari, M., Jalali, M., Parras, S.J., Bayer, P., 2017. Synthetic fracture network characterization with transdimensional inversion. Water Resour. Res. 53, 5104–5123. https://doi.org/10.1002/2016WR020293.Received

Stanford, K., 2017. Underdetermination of Scientific Theory, in: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University.

Strebelle, S., 2002. Conditional Simulation of Complex Geological Structures Using Multiple-Point Statistics. Math. Geol. 34, 1–21. https://doi.org/10.1023/A:1014009426274

Sugiura, N., 1978. Further analysts of the data by Akaike’ s information criterion and the finite corrections. Commun. Stat. - Theory Methods 7, 13–26. https://doi.org/10.1080/03610927808827599

Sun, N.-Z., Yeh, W.W.G., 1985. Identification of Parameter Structure in Groundwater Inverse Problem. Water Resour. Res. 21, 869–883.

Suzuki, S., Caumon, G., Caers, J., 2008. Dynamic data integration for structural modeling: Model screening approach using a distance-based model parameterization. Comput. Geosci. 12, 105–119. https://doi.org/10.1007/s10596-007-9063-9

Tarantola, A., 2006. Popper, Bayes and the inverse problem. Nat. Phys. 2, 4–7.

Telford, W.M., Telford, W.M., 1976. Applied geophysics. Cambridge University Press, London; New York.

Terrell, B.L., Johnson, P.N., Segarra, E., 2002. Ogallala aquifer depletion: economic impact on the Texas high plains. Water Policy 4, 33–46.

Thompson, S., MacVean, L., Sivapalan, M., 2017. A stochastic water balance framework for lowland watersheds. Water Resour. Res. 53, 9564–9579. https://doi.org/10.1002/2017WR021193

Thomsen, N.I., Binning, P.J., Mcknight, U.S., Tuxen, N., Bjerg, P.L., Troldborg, M., 2016. A Bayesian belief network approach for assessing uncertainty in conceptual site models at contaminated sites. J. Contam. Hydrol. 188, 12–28. https://doi.org/10.1016/j.jconhyd.2016.02.003

Tickell, S.J., 2013. A recent sinkhole collapse at Lambells Lagoon. North. Territ. Dep. L. Resour. Manag. Tech. Rep. No. 02/2013D.

Tickell, S.J., Zaar, U., 2017. Water Resources of the Wildman River area (Technical Report 8/2017D). Northern Territory Department of Environment and Natural Resources.

Tonkin, M., Doherty, J., Moore, C., 2007. Efficient nonlinear predictive error variance for highly parameterized models. Water Resour. Res. 43, 1–15. https://doi.org/10.1029/2006WR005348

Troldborg, L., Refsgaard, J.C., Jensen, K.H., Engesgaard, P., 2007. The importance of alternative conceptual models for simulation of concentrations in a multi-aquifer system. Hydrogeol. J. 15, 843–860. https://doi.org/10.1007/s10040-007-0192-y

Troldborg, M., Nowak, W., Tuxen, N., Bjerg, P.L., Helmig, R., Binning, P.J., 2010.

Uncertainty evaluation of mass discharge estimates from a contaminated site using a fully Bayesian framework. Water Resour. Res. 46, 1–19. https://doi.org/10.1029/2010WR009227

Tsai, F.T.C., 2010. Bayesian model averaging assessment on groundwater management under model structure uncertainty. Stoch. Environ. Res. Risk Assess. 24, 845–861. https://doi.org/10.1007/s00477-010-0382-3

Tsai, F.T.C., Elshall, A.S., 2013. Hierarchical Bayesian model averaging for hydrostratigraphic modeling: Uncertainty segregation and comparative evaluation. Water Resour. Res. 49, 5520–5536. https://doi.org/10.1002/wrcr.20428

Tsai, F.T.C., Li, X., 2008. Multiple parameterization for hydraulic conductivity identification. Ground Water 46, 851–864. https://doi.org/10.1111/j.1745-6584.2008.00478.x

Tsang, C., 1991. The Modelling Process and Model Validation. Ground Water 29, 825–831.

Tsang, C., 1987. Technical Note: Comments on Model Validation. Transp. Porous Media 2, 623–629.

Turnadge, C., Crosbie, R.S., Tickell, S.J., Zaar, U., Smith, S.D., Dawes, W.R., Davies, P., Harrington, G.A., Taylor, A.R., 2018a. Hydrogeological characterisation of the Mary–Wildman rivers area, Northern Territory, in: A Technical Report to the Australian Government from the CSIRO Northern Australia Water Resource Assessment, Part of the National Water Infrastructure Development Fund: Water Resource Assessments.

Turnadge, C., Mallants, D., Peeters, L.J.M., 2018b. Sensitivity and uncertainty analysis of a regional-scale groundwater flow model featuring coal seam gas extraction.

Turnadge, C., Taylor, A.R., Harrington, G.A., 2018c. Groundwater flow modelling of the Mary – Wildman rivers area, Northern Territory., in: A Technical Report to the

Australian Government from the CSIRO Northern Australia Water Resource Assessment, Part of the National Water Infrastructure Development Fund: Water Resources Assessments.

Usunoff, E., Carrera, J., Mousavi, S.F., 1992. An approach to the design of experiments for discriminating among alternative conceptual models. Adv. Water Resour. 15, 199–214. https://doi.org/10.1016/0309-1708(92)90024-V

Vanorio, T., Prasad, M., Nur, A., 2003. Elastic properties of dry clay mineral aggregates, suspensions and sandstones. Geophys. J. Int. 155, 319–326. https://doi.org/10.1046/j.1365-246X.2003.02046.x

Vrugt, J.A., 2016. Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation. Environ. Model. Softw. 75, 273–316. https://doi.org/10.1016/j.envsoft.2015.08.013

Vrugt, J.A., Robinson, B.A., 2007. Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging. Water Resour. Res. 43. https://doi.org/10.1029/2005WR004838

Walker, W.E., Harremoes, P., Rotmans, J., van der Sluijs, J.P., van Asselt, M.B.A., Janssen, P., Krayer Von Krauss, M.P., 2003. A Conceptual Basis for Uncertainty Management. Integr. Assesment 4.

Wang, Z. (Zee), Wang, H., Cates, M.E., 2001. Effective elastic properties of solid clays. Geophysics 66, 428–440. https://doi.org/10.1190/1.1444934

Webb, G.I., 2017. Bayes' Rule, in: Sammut, C., Webb, G.I. (Eds.), Encyclopedia of Machine Learning and Data Mining. Springer, Boston, MA. https://doi.org/10.1007/978-1-4899-7687-1

White, J.T., Doherty, J.E., Hughes, J.D., 2014. Quantifying the predictive consequences of model error with linear subspace analysis. Water Resour. Res. 50, 1152–1173. https://doi.org/10.1002/2013WR014767

Wingefors, S., Andersson, J., Norrby, S., Eisenberg, N.A., Lee, M.P., Federline, M.V., Sagar, B., Wittmeyer, G.W., 1999. Regulatory perspectives on model validation in high-level radioactive waste management programs: A Joint NRC/SKI White Paper. Stockholm, Sweden and Washington DC, USA.

Winter, C.L., Nychka, D., 2010. Forecasting skill of model averages. Stoch. Environ. Res. Risk Assess. 24, 633–638. https://doi.org/10.1007/s00477-009-0350-y

Winter, T.C., Harvey, J.W., Franke, O.L., Alley, W.M., 1998. Groundwater and surface water: A single resource, U.S. Geological Survey Circular 1139. Denver, Colorado.

Woolfenden, L.R., 2008. Use of a groundwater flow model to assess the location, extent , and hydrologic properties of faults in the Rialto-Colton Basin, California, in: MODFLOW and More 2008. pp. 78–82.

Yakirevich, A., Pachepsky, Y.A., Gish, T.J., Guber, A.K., Kuznetsov, M.Y., Cady, R.E., Nicholson, T.J., 2013. Augmentation of groundwater monitoring networks using information theory and ensemble modeling with pedotransfer functions. J. Hydrol. 501, 13–24. https://doi.org/10.1016/j.jhydrol.2013.07.032

Ye, M., Meyer, P.D., Neuman, S.P., 2008a. On model selection criteria in multimodel analysis. Water Resour. Res. 44, 1–12. https://doi.org/10.1029/2008WR006803

Ye, M., Neuman, S.P., Meyer, P.D., 2004. Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff. Water Resour. Res. 40, 1–17. https://doi.org/10.1029/2003WR002557

Ye, M., Neuman, S.P., Meyer, P.D., Pohlmann, K.F., 2005. Sensitivity analysis and assessment of prior model probabilities in MLBMA with application to unsaturated fractured tuff. Water Resour. Res. 41, 1–14. https://doi.org/10.1029/2005WR004260

Ye, M., Pohlmann, K.F., Chapman, J.B., 2008b. Expert elicitation of recharge model probabilities for the Death Valley regional flow system. J. Hydrol. 354, 102–115. https://doi.org/10.1016/j.jhydrol.2008.03.001

Ye, M., Pohlmann, K.F., Chapman, J.B., Pohll, G.M., Reeves, D.M., 2010. A Model-Averaging Method for Assessing Groundwater Conceptual Model Uncertainty. Groundwater 48, 716–728. https://doi.org/10.1111/j.1745-6584.2009.00633.x

Young, P., Parkinson, S., Lees, M., 1996. Simplicity out of complexity in environmental modelling: Occam's razor revisited, Journal of Applied Statistics. https://doi.org/10.1080/02664769624206

Youssef, A.M., El-Kaliouby, H.M., Zabramawi, Y.A., 2012. Integration of remote sensing and electrical resistivity methods in sinkhole investigation in Saudi Arabia. J. Appl. Geophys. 87, 28–39. https://doi.org/10.1016/j.jappgeo.2012.09.001

Zelt, C.A., Haines, S., Powers, M.H., Sheehan, J., Rohdewald, S., Link, C., Hayashi, K., Zhao, D., Zhou, H.W., Burton, B.L., Petersen, U.K., Bonal, N.D., Doll, W.E., 2013. Blind test of methods for obtaining 2-d near-surface seismic velocity models from first-arrival traveltimes. J. Environ. Eng. Geophys. 18, 183–194. https://doi.org/10.2113/JEEG18.3.183

Zeng, X., Wang, D., Wu, J., Zhu, X., Wang, L., Zou, X., 2015. Evaluation of a Groundwater Conceptual Model by Using a Multimodel Averaging Method. Hum. Ecol. Risk Assess. An Int. J. 21, 1246–1258. https://doi.org/10.1080/10807039.2014.957945

Zhou, Y., Herath, H.M.P.S.D., 2016. Evaluation of alternative conceptual models for groundwater modelling. Geosci. Front. 8, 437–443. https://doi.org/10.1016/j.gsf.2016.02.002

Zomlot, Z., Verbeiren, B., Huysmans, M., Batelaan, O., 2017. Trajectory analysis of land use and land cover maps to improve spatial–temporal patterns, and impact assessment on groundwater recharge. J. Hydrol. 554, 558–569. https://doi.org/10.1016/j.jhydrol.2017.09.032

Zyvoloski, G., Kwicklis, E., Eddebbarh, A.A., Arnold, B., Faunt, C., Robinson, B.A., 2003. The site-scale saturated zone flow model for Yucca Mountain: Calibration of different conceptual models and their impact on flow paths. J. Contam. Hydrol. 62–63, 731–750. https://doi.org/10.1016/S0169-7722(02)00190-0