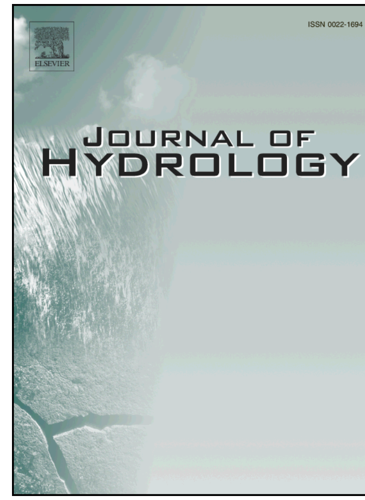
Flinders UNIVERSITY



Hydrogeological conceptual model building and testing: A review

Trine Enemark, Luk J.M. Peeters, Dirk Mallants, Okke Batelaan

PII: S0022-1694(18)30938-7
DOI: https://doi.org/10.1016/j.jhydrol.2018.12.007
Reference: HYDROL 23316

To appear in: *Journal of Hydrology*

Received Date: 27 July 2018
Revised Date: 6 November 2018
Accepted Date: 4 December 2018

Hydrogeological conceptual model building and testing: A review

Trine Enemark [a,b] (corresponding author); trine.enemark@csiro.au

Luk JM Peeters [a]; luk.peeters@csiro.au

Dirk Mallants [a]; dirk.mallants@csiro.au

Okke Batelaan [b]; okke.batelaan@flinders.edu.au

[a] CSIRO Land and Water, Gate 4 Waite Rd, Locked Bag 2, Glen Osmond SA 5064 Australia;

[b] National Centre for Groundwater Research and Training, College of Science & Engineering, Flinders University, Adelaide, SA 5001, Australia

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.

Keywords

Conceptual models; model evaluation; model rejection; multi-model framework; conceptual model uncertainty.

1 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) (Fig. 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.

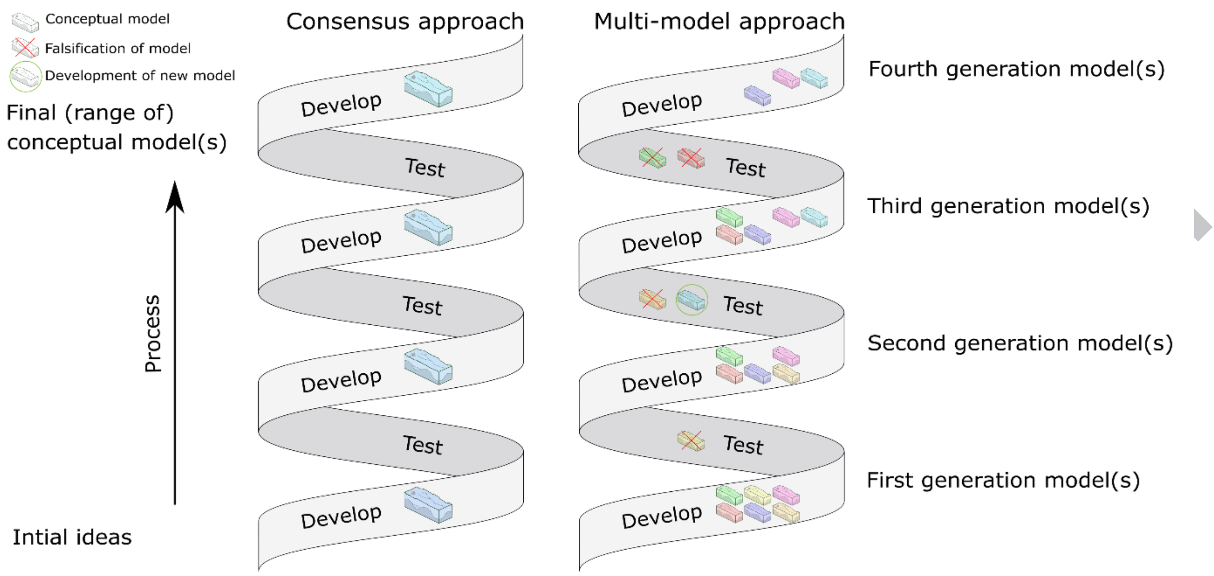


Fig. 1. Iterative process for the conceptual modelling process via the consensus or multi-model approach. Modified from Environment Agency (2002) and Suzuki et al. (2008). Each model test step involves introducing new data and thereby identifying new plausible models uncovering conceptual surprises, 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 Fig. 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 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 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.

2.1 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 (Fig. 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.

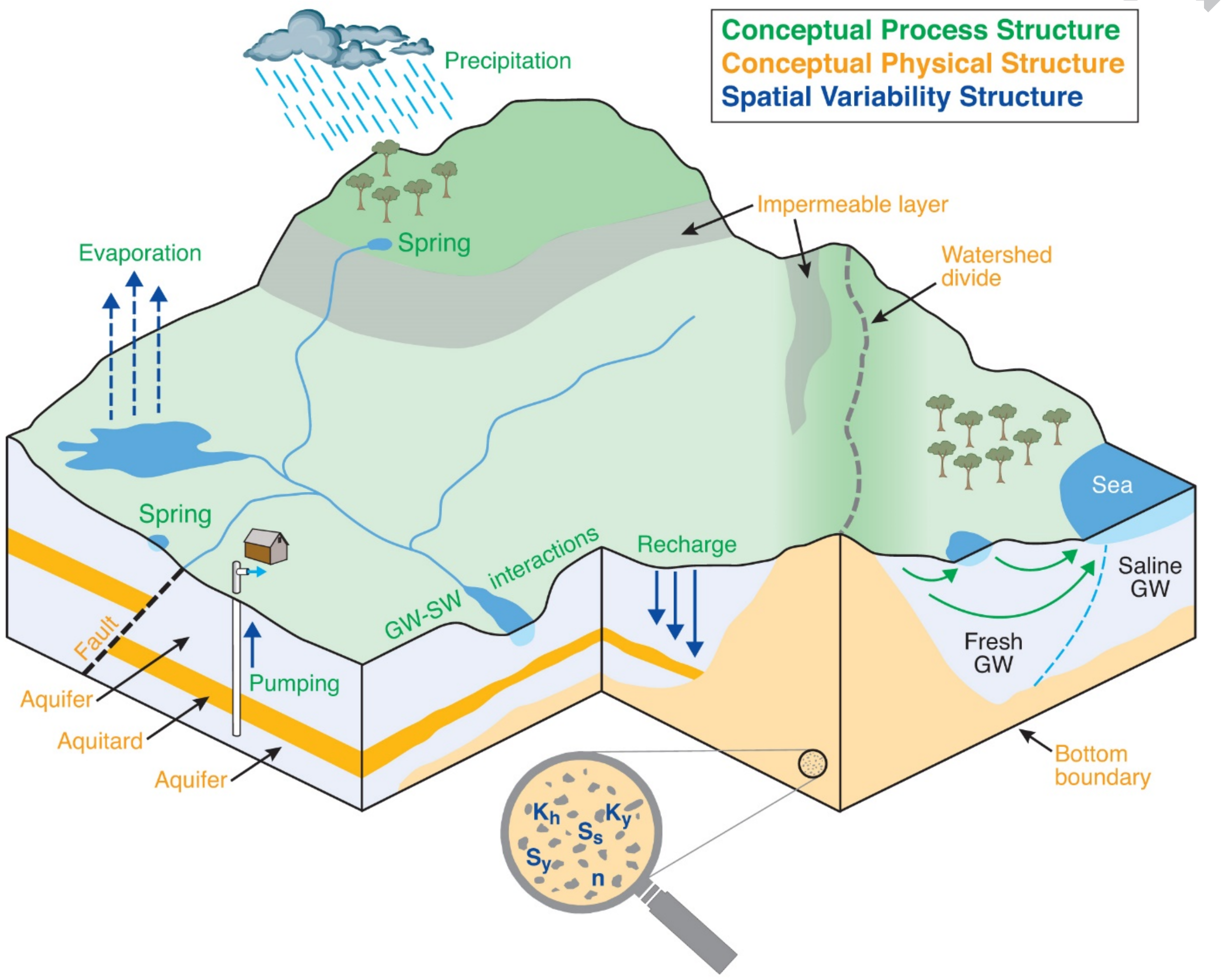


Fig. 2. Elements of a conceptual model. Items in green illustrate the Conceptual Process Structure, while items in blue illustrate the Spatial Variability Structure represented in the magnifying glass (K_h = horizontal hydraulic conductivity, K_v = vertical hydraulic conductivity, n=porosity, S_s = Specific 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 Fig. 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 Fig. 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).

2.2 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, 2017). 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 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.

2.4 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.

3 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 Fig. 3.

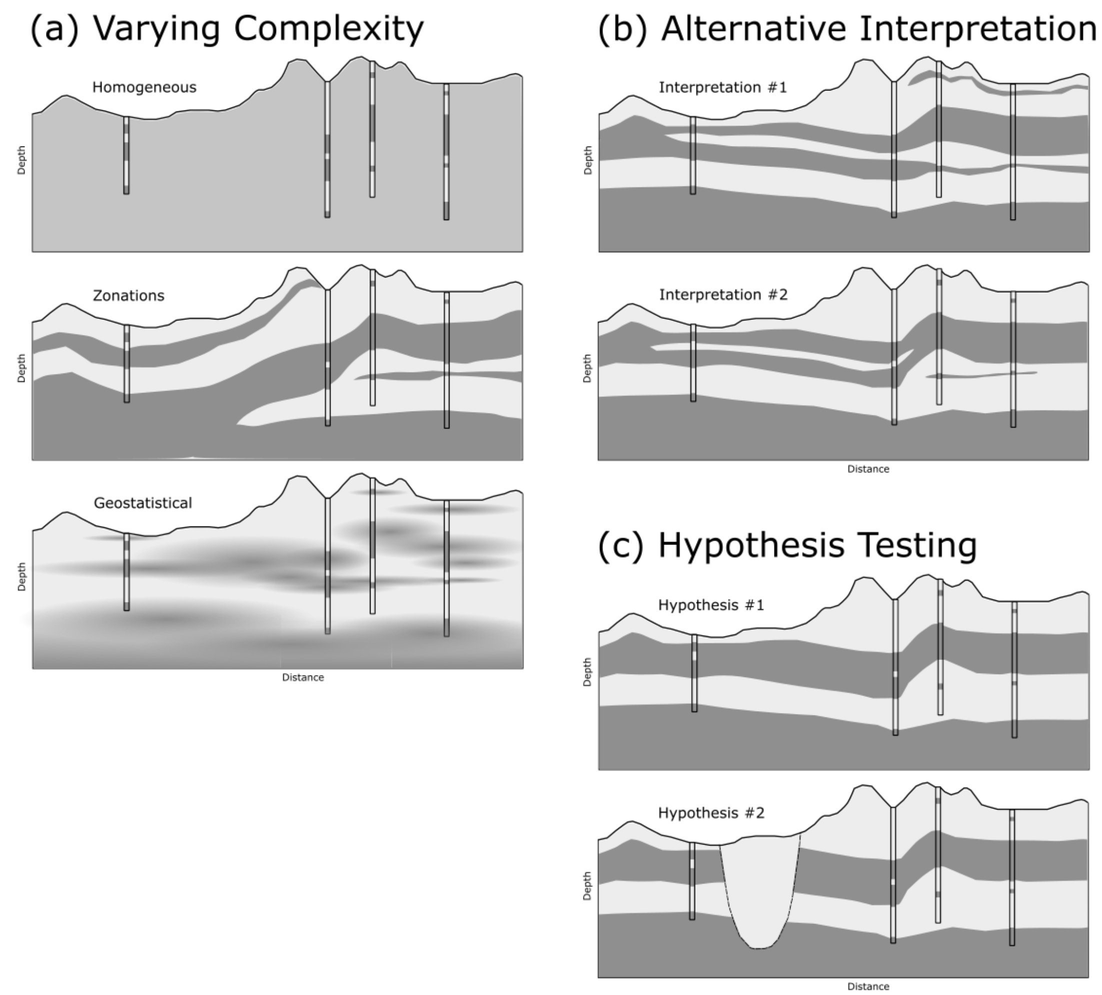


Fig. 3. Conceptual model development approaches in the multi-model approach. Illustration of how different conceptualizations of the Conceptual Physical Structure could take shape if based on the same data (boreholes in this case) through Varying Complexity (a), Alternative Interpretation (b) or Hypothesis 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 Fig. 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., 2015), 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. Fig. 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 Fig. 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 Fig. 2; Conceptual Physical Structure (section 3.1), Spatial Variability Structure (section 3.2), and Conceptual Process Structure (section 3.3). 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 (Fig. 3). Finally, section 3.4 discusses assigning prior probabilities to alternative models.

3.1 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 regards to that specific feature.

3.2 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

3.3 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.

3.4 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 4). 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 Fig. 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.

3.5 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 have to 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, 2005).

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 hypothesis 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 Fig. 1, is a part of the modelling plan. The second method anticipates surprises by placing fewer restrictions on what is considered possible. By 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 were 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, taking into account the largest possible range of the conceptual uncertainty, have not been applied yet and remain a subject for further research.

4 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 illustrate 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 5.1) while only eight studies apply a model testing procedure.

In the subsequent sections, data relevant to conceptual model testing (section 4.1), steps undertaken when testing conceptual models (section 4.2), and the remaining challenges within model testing (4.3) are discussed. Table A.2 presents an overview of the model testing applied in the studies identified using the multi-model approach (Section 2).

4.1 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).

4.1.1 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.

4.1.2 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).

4.1.3 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 actually 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.

4.2 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 (**Error! Reference source not found.**). In addition to these three steps, a fourth step, the post-audit (Anderson and Woessner, 1992) will be reviewed here.

Table 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 is able to 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)	Nordqvist and Voss (1996) concluded that a supply well was in risk of contamination through a multi-model approach. After the

	Parameters can compensate for conceptual error (C)	completion of the study, increased levels of contamination were observed in the well which conditionally validated the models.

4.2.1 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).

4.2.2 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.

4.2.3 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., 2015; 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.

4.2.4 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 4.1). 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.

4.3 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 5.1), 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 actually 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 4.2, 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 in itself 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.

5 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 4.1), while other studies (e.g. Tsai and Li, 2008) focus on ensemble predictions based on all plausible models (section 4.2). 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 4.3). 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 3) affects the predictions.

5.1 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 5.2). 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.

5.2 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.

5.3 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 4.2) and the use of these data for model testing will ensure the data is independent from the model development data.

5.4 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 3) 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 4.1, 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 taken into account. 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. 5^{th}, 50^{th} and 95^{th}) 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 3.5. 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. Transdimensional 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.

6 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 recommend, 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 regards to both developing and testing alternative models.

7 Acknowledgement

The authors would like to thank Russell Crosbie, Rebecca Doble, Hoshin V. Gupta, Ty Ferre, Rodrigo Rojas and one anonymous reviewer for review and constructive comments on the manuscript. This research was conducted as part of a PhD project funded by CSIRO.

8 Appendix A

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 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). Streams are described with MODFLOW's SFR and River package in alternative models imposing different assumptions (H).	I	C	I/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 3 therefore does not apply.	N/A	N/A	N/A

Study	Is the model objective well defined?	Conceptual multi-model development approach	Ph	Pr	SVS
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
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.		C	C

Study	Is the model objective well defined?	Conceptual multi-model development approach	Ph	Pr	SVS
		A steady-state and a transient boundary condition to a stream.			
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). 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).		C	C/H
Samper and Neuman (1989)	No	Five different semi variogram models (exponential, quadratic, spherical, pure nugget and exponential with nugget).		H	

Study	Is the model objective well defined?	Conceptual multi-model development approach	Ph	Pr	SVS
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		
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		

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			
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	-	-

Study	Prior	Model Testing			Model Predictions			
	Uninformed/ informed	Step 1	Step 2	Step 3	Model Ranking	Individual Predictions	Ensemble Predictions	Additional data needs
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	-	-
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.	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
				Model convergence.				
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	-	-
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]

Study	**Prior**	**Model Testing**			**Model Predictions**			
	Uninformed/ informed	**Step 1**	**Step 2**	**Step 3**	**Model Ranking**	**Individual Predictions**	**Ensemble Predictions**	**Additional data needs**
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).

9 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

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

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., 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

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

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., 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

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

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

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.

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

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

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

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

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

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.

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., 2005. Analytic element modeling of groundwater flow. Academic Press.

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

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

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

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., 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.

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

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

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

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

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

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, 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., 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

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

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

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

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., 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.

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

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.

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

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

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. Researh 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

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

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

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

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

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., 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

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., 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

Schöniger, A., Illman, W.A., Wöhling, T., Nowak, W., 2015. 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., 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

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

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., 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

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

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.

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.

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

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.

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

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., 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., 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

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., 2017. Evaluation of alternative conceptual models for groundwater modelling. Geosci. Front. 8, 437–443. https://doi.org/10.1016/j.gsf.2016.02.002

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

Hydrogeological conceptual model building and testing: A review

Highlights

- Reviewed 59 studies that applied hydrogeological multi-model approach.
- Developing mutually exclusive, collectively exhaustive models remains a challenge.
- Conceptual model testing is underutilised but can uncover inconsistent assumptions.
- Iterative model development and testing accommodate conceptual “surprises”.
- Model testing is limited by the independence and information content of data.