A Systematic Approach to Hydrogeological Conceptual Model Testing, Combining Remote Sensing and Geophysical Data

Trine Enemark1,2, Luk Peeters1, Dirk Mallants1, Brady Flinchum1, and Okke Batelaan2

1Land and Water, CSIRO, Glen Osmond, South Australia, Australia, 2National Centre for Groundwater Research and Training College of Science and Engineering, Flinders University, Bedford Park, South Australia, Australia

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

1. Introduction

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

In this context, the multimodel approach presents a promising tool to increase transparency, reproducibility, and to integrate an automation of an expert’s thought in the modeling workflow (Enemark et al., 2019). In the multimodel approach, alternative conceptual models are proposed and evaluated. The multimodel approach allows for a transparent account of model choices, potential rejection of invalid conceptual models, and unveiling of conceptual “surprises” (Ferré, 2017). The approach is increasingly applied to explore conceptual uncertainty in hydrogeology (Mustafa et al., 2020; Timani & Peralta, 2015) and hydrology (Clark et al., 2008; Duan et al., 2007).

A natural extension of developing an ensemble of alternative conceptual models is to choose the best performing model among them. Approaches to choose between alternative conceptual models were classified into three broad categories by Carrerra et al. (1993): (1) comparative analysis of predicted and observed values (Pirot et al., 2015; Zeng et al., 2015), (2) assessment of calibrated parameter values to observed values (Engelhardt et al., 2014; Poeter & Anderson, 2005), and (3) “identification criteria” (model selection...
criteria), which are often based on maximum likelihood and may also consider the principle of parsimony (Schöniger et al., 2014).

Removing alternative models from the model ensemble in this way requires that another model exists that outperforms the removed model (Gelman & Shalizi, 2013), and it is therefore implicitly assumed that the range of alternative models is collectively exhaustive (Enemark, Peeters, Mallants, & Batelaan, 2019). Further, the best performing model becomes unfalsifiable, meaning that conceptual “surprises” cannot be uncovered even if the model that most adequately represents the real system has not been developed yet.

The Popper-Bayes philosophy, which is hypothesis testing applying a Bayesian framework combined with a falsification approach (Gelman & Shalizi, 2013; Linde et al., 2015; Tarantola, 2006), offers an alternative framework to choose between conceptual models. This hypothetico-deductive framework builds on the idea that “observations cannot produce models; they can only falsify models.” The falsification type approach consists of checking the models against data and rejecting models that are inconsistent. The Bayesian framework on the other hand offers a systematic approach to updating the prior beliefs about the adequacy of a model to posterior beliefs.

The framework consists of rejecting and updating the probability of alternative conceptual models through forward modeling. Traditionally, the difference between a forward modeled response and the observed response is used to drive an inversion to find the model that best explains data. Usually the model with the lowest mismatch between observed and modeled response is considered the best model. In the Popper-Bayes methodology, instead of relying on the inversions which are prone to spatial averaging and smoothing to interpret the conceptual model, the forward modeled response is used to justify an ensemble of conceptual understandings that can be sufficiently explained by observed data.

A Popper-Bayes framework has been applied by researchers in various fields of water resources (Hermans et al., 2015; Park et al., 2013; Pirot et al., 2019; Scheidt et al., 2015). The framework is however still largely an academic exercise, prompting the delineation of a workflow that can readily be adopted by practitioners. We will therefore delineate a Popper-Bayes workflow by unifying current insights into individual steps that can be generally applied to a model testing exercise. Also, in the currently available applications, a single data type has been used to test alternative models. Oreskes et al. (1994) argued that “the greater the number and diversity of confirming observations, the more probable it is that the conceptualization embodied in the [forward] model is not flawed.” When the objective of the testing exercise is to explore the model space, this holds especially true. In the workflow applied here, we will therefore allow for multiple data types to be used for testing. Finally, in the workflow, the model rejection step will be clearly separated from updating the probability. This is to avoid assuming that the model ensemble is collectively exhaustive and ensure all models can potentially be rejected by data.

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

2. Methodology

The systematic testing approach for exploratory analysis of conceptual models is presented in Figure 1. A hydrogeological conceptual model is a summary of the current knowledge about the groundwater system, and it consists of a process structure and a physical structure (Carrera et al., 1993; Gupta et al., 2012). The physical structure describes the hydrostratigraphy and geometry of aquifers, while the process structure contains the internally and externally controlled boundary conditions such as heads and fluxes in a system.

The workflow starts by developing alternative models of a prior conceptualization representing different understandings of groundwater system functioning (step 1). Data independent of the model development data, that can test the alternative conceptualizations, are then identified (step 2). The qualitative conceptual models are parameterized in forward models that provide an output that can be compared to the independent data (step 3). The correspondence between multiple model realizations and observations can then be
used to reject the unsupported models from the prior model ensemble (step 4). Further, the probability of the remaining alternative understandings can be updated (step 5). If all alternative models are rejected or if new data for model testing are available, the workflow can be repeated (dashed lines in Figure 1).

In the following sections, we provide an overview of the individual components in the model testing workflow (Figure 1). This systematic conceptual model testing workflow is subsequently demonstrated on a case study area in the Wildman River area, Australia, where sinkhole-type depressions are a common feature of the landscape, potentially contributing to highly localized groundwater recharge.

2.1. Model Development

Developing alternative conceptual models presents a natural first step of the systematic testing exercise (Figure 1). Alternative understandings can be based on a literature review of the site under investigation as well as general or subject-matter knowledge about groundwater system functioning (Chatfield, 1995; Clark et al., 2011). During the site literature review, one must identify the rationale underpinning each assumption to ascertain whether alternative understandings could be possible or not (Peeters, 2017).

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

2.2. Data Identification

The second step in the testing workflow involves identifying independent observation data (Figure 1). The requirement for the testing data is that it should be independent from the model development data (i.e., data used in step 1, Figure 1) in order to avoid circular reasoning, overconfidence in the conceptual models and under-sampling of the model space (Chatfield, 1995; Kerr, 1998). Without explicitly reserving data for model testing, the remaining available data, independent of the model development, might have little power to test the model (Rojas et al., 2010). The power of the model testing data relates to the type, amount, and uncertainty of that data. The consequence of testing alternative model conceptualizations with data that have limited information content is that discrimination among alternative models cannot be made (Enemark et al., 2020).
et al., 2019; Seifert et al., 2008) and, in a Bayesian context, that the prior probability has a large influence on the posterior probability (Rojas et al., 2009).

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

### 2.3. Forward Modeling

The third step in the conceptual model testing workflow (Figure 1) involves translating the qualitative alternative conceptual models, defined in the first step, into forward models. A forward model is a simplified mathematical description of a system that captures the main physical or process structures, predicting a response of a given conceptual model that can be compared to observations.

In hydrogeology the forward model is most often a numerical groundwater model (Remson et al., 1980) or solute transport model (Thomsen et al., 2016; Trolldborg et al., 2010), but the forward model can also be a geophysical forward model (Brunetti et al., 2017; Hermans et al., 2015) or a water balance model (Enemark, Peeters, Mallants, Batelaan, Valentine, & Sambridge, 2019).

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

### 2.4. Model Rejection

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

Conceptual models are rejected based on the assumption that success and failures of model realizations relate to the model structure and not, for example, to undetected uncertainties in the observation data or incorrectly defined parameter values. In case all alternative models can be rejected, a conceptual “surprise” (Bredehoeft, 2005) has been uncovered and the real model structure must be considered to be an unknown. This means one should start over and develop new conceptualizations (step 1, Figure 1).

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

### 2.5. Model Probability Update

The final step in the model testing workflow (Figure 1) is updating the prior probability of the remaining alternative models (after the model rejection step) to a posterior probability.

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

$$p(m_k | Y_{pl, ps}) = \frac{p(Y_k | m_k) \cdot p(m_k)}{\sum_{i=1}^{N_m} p(Y_k | m_i) \cdot p(m_i)}$$

(1)

where the conceptual model element $m_k = [m_{pl}, m_{ps}]$ and the observed data $Y_k = [Y_{pl}, Y_{ps}]$ refer to either the physical structure ($pl$) or the process structure ($ps$), that are both part of the conceptual model $m_c$. The denominator, where $N_m$ is the number of defined, plausible conceptual models, normalizes the sum of posterior probabilities to 1.
The prior probability of each model structure \( p(m_k) \) describes our belief in each model before any data are considered. The prior model probability is therefore defined in the model development phase (step 1, Figure 1). If we can assume that the physical structure influences the process structure, then the priors can be defined as

\[
p(m_k) = p(m_{pl}) \cdot p(m_{pl}|m_{ps})
\]

(2)

The observed data types are independent of each other, if the physical structure testing data set does not depend on any value in the process structure testing data set and vice versa. If the data types are independent, the model evidence \( p(Y_k|m_k) \) describing the probability of the observed data over each model’s parameter space is

\[
p(Y_k|m_k) = p(Y_{pl}|m_{pl}) \cdot p(Y_{ps}|m_{ps})
\]

(3)

The model evidence and thereby also the posterior probability rely on the relative performance of the individual model structures of the ensemble against data. To evaluate the model evidence of the individual models, we sample each model’s parameter space from prior parameter probabilities. As indicated in Figure 1, if new data exist/are collected, the conceptual model probabilities can be updated again.

### 3. Methodology Implementation

In this section, we implement the methodology discussed in section 2 to a simplified hydrogeological characterization of the Wildman River area (Northern Territory, Australia) and evaluate the connectivity of sinkhole-type depressions to groundwater. First, a brief description of the depressions is provided in section 3.1. In sections 3.2–3.6, the individual steps in the suggested workflow (Figure 1) are discussed as they are applied to the case study area.

#### 3.1. Depressions in the Wildman River Area

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

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

In the Wildman River area the dolostone is known to be less continuous but similar depressions are found. A x-ray diffraction analysis found that the dolostone mineralogical composition is mainly dolomite with fractions of muscovite and quartz (Turnadge et al., 2018) suggesting that dissolution and sinkhole formation may also have taken place. The locations of possible sinkhole features in the Wildman River area have been mapped by several sources (Easey et al., 2016; Mueller et al., 2016; Tickell & Zaar, 2017; Figure 2) and are described further in Supporting Information S1.

At the end of the dry season (late October 2018), a fieldtrip was undertaken to investigate a subset of the mapped depressions in the Wildman River area. Five depressions (S1 to S5, Figure 2) were selected based on their accessibility and vicinity to existing boreholes and water level loggers. We prioritized the selection of sinkholes based on their difference in geometry and vegetation cover while making sure they covered different parts of the area. The collected data at each of the five depressions included a refraction seismic line, high water level marks, and topography. Further, PlanetScope satellite imagery (Planet Team, 2017) has been collected over the area.
3.2. Model Development

Previous investigations have hypothesized depressions in Wildman River area to be sinkholes that may act as conduits for recharge (Graham, 1985; Turnadge et al., 2018). This corresponds to observations that water levels in surficial aquifers respond within days of major rainfall events (Tickell & Zaar, 2017). Whether these sinkhole-like depressions act as conduits of recharge is an important conceptual question to resolve as it creates a potential for contamination to rapidly reach the water resource. The null and alternative hypotheses, respectively, $H_0$ and $H_A$, tested in this paper are defined as follows:

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

The conceptual understanding includes a process structure and physical structure (Figure 3). The physical structure accounts for different lithologies underlying the depressions (i.e., sand or clay), while process structure accounts for different interactions (i.e., fluxes) between the surface water and the groundwater. As stated in Figure 1, the alternative model structures are based on a literature review.

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

Lakes interact with groundwater in three basic ways (Winter et al., 1998): (1) the lake receives groundwater discharge throughout the entire lakebed, (2) the lake loses groundwater recharge throughout the entire lakebed, and (3) part of the lake bed receives groundwater discharge, while other parts of the lakebed lose groundwater recharge. In addition to these three modes of interaction, the depressions may not interact with groundwater at all. In case 2, the water level in the depression may be connected or disconnected from the water table; however, the difference between these two will not be studied here. In the four developed process structure models, the water balance of the depression is influenced by the following:

- I: Evaporation only, there is no exchange between surface and groundwater.
- II: Evaporation and groundwater recharge, where the groundwater flux occurs under either losing-connected or losing-disconnected conditions.
- III: Evaporation and groundwater discharge.
- IV: Evaporation, groundwater recharge, and groundwater discharge.
Note that the influence of surface water input or output is not considered as the depressions are enclosed features. It should also be noted that due to the available independent data (introduced in section 3.3), the process structure models will only be tested for the dry season as a single long-term event.

The physical structures \((m_{pl})\) are based on an understanding of the geomorphological development of depressions. Turnadge et al. (2018) suggested the depressions have formed as either dropout or buried sink-holes that develop on top of areas of relatively higher porosity where surface water infiltrates preferentially. The infiltrating water eventually results in dissolution of the dolomitic bedrock and forms underground cavities that ultimately lead to the collapse of the overlying features. This development history suggests that the permeability underneath the sinkhole could be greater than outside the sinkholes. Other observations indicate sinkholes could evolve into shallow depressions (Tickell, 2013) that over time develop a layer of low permeability clay and organic matter effectively sealing it from the groundwater (Graham, 1985; Schult & Welch, 2006). Combination of the understandings of the development history of the depressions leads to four different physical structure models:

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

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

### 3.2.1. Prior Probabilities \(p(m_{pl}), p(m_{ps})\)

Current knowledge of the study area does not suggest any of the proposed model structures being more probable than others. However, the model structures are combined to \((4 \times 4 = 16\) different conceptual models, and as the processes are mediated by the physical structure (Gupta et al., 2012), we assume that the specified physical structure influences the probability of the specified process structure. The groundwater exchange in the process structures (II, III, and IV) combined with a sealing clay layer in the physical structures (C and D)

---

**Figure 1.** Alternative conceptual understandings of the physical and process structure of depressions in the Wildman River area. The alternative understandings of the physical structure concern the presence (C and D) or absence (A and B) of a clay layer and the presence (B and D) or absence (A and C) of a high-porosity zone in the middle of the depression. The dashed line in the physical structures represents an initial topography of the depression. The alternative understandings of the process structure consider whether the depression interacts (II, III, and IV) or does not interact with the groundwater (I). Groundwater interactions tested are groundwater recharge (II), groundwater discharge (III), and whether the depressions are flow-through features with both groundwater discharge and recharge (IV). In structure II we do not attempt to differentiate whether the groundwater recharge feature is directly connected or disconnected to the groundwater.
is thereby assumed only half as probable as other combinations of the structures. This expresses our understanding of the physics of groundwater flow and the hydraulic barrier effect of clay.

Assuming an otherwise uniform prior probability and that the prior probabilities add up to one, models including a sealing clay layer combined with any groundwater exchange are assigned a probability of 0.0385, while all other contending models are assigned a prior probability of 0.0769 (Table 1).

Note that the prior probability assignment is subjective and the effect of these choices on the posterior probability could be evaluated by exploring the result of other prior probabilities. This is however outside the scope of this study. It should also be noted that by assuming probabilities add up to one, we are implicitly assuming that the range of models is collectively exhaustive, which is probably not the case. However, any remaining conceptual models are unknown unknowns.
steps ($t$) over the dry season. The depth at a given time step $t$ depends on the depth at time step $t-1$ and on inflows and outflows:

$$d_t = d_{t-1} - \text{inflow} + \text{outflow}$$  \hspace{1cm} (4)$$

The result from the water balance calculations of interest here is the time step when the depth of the water in the depression is 0. The input to the forward models consists of samples from prior parameter distributions of maximum water depth in the individual depressions and rates of evaporation, groundwater recharge, and groundwater discharge.

The forward response from the alternative physical structures is obtained from a combination of a rock physics modeling and ray tracing using a shortest path algorithm (Moser, 1991; Rücker et al., 2017). Rock physics relationships are utilized to provide reasonable estimates for the velocities expected by the lithologies in the physical structure. Following the methods of Flinchum et al. (2018) and Holbrook et al. (2014), we estimate the seismic velocity of a porous media by using a Hashin-Shtrikman relationship to define velocities as a function of porosity and then use Gassmann’s equations (Brie et al., 1995; Mavko et al., 2009) to adjust for water saturation. The rock physics model is further described in Supporting Information S3.

The velocities from the rock physics model and the geometry from the physical structure are projected onto a triangular mesh and passed to the open-source Python framework pyGIMLi (Rücker et al., 2017) to compute the first arrival travel-times (Figures 4a–4d). By using the same configuration as for the collection of field data, the travel time from each source to each receiver is calculated and can be directly compared to the travel-times observed in the field.

3.5. Model Rejection

In the model rejection step (step 4, Figure 1), the two independent data sets (remote sensing and seismic refraction) are used to reject the alternative models of process structure and physical structure.

The likelihood function of the last day of water inundation based on the remote sensing data (Section 3.3) has a bounded interval describing possible values. This means realizations from the forward model that lie outside of the bounds are assigned a zero likelihood. Model structures obtaining a zero marginal likelihood are rejected; that is, the model evidence $p(Y_{ps} | m_{ps})$ is set to 0.

To globally test the physical models, we apply a dimension reduction technique to the refraction seismic data. We apply a principal component analysis (PCA) using the PCA class from the open-source Python framework scikit-learn (Pedregosa et al., 2011) that reduces the number of dimensions in a dataset by feature extraction. From the first arrival variables (>900 dimensions), a smaller number of new variables (principal components) are created that explain the variability of the old variable (Figures 4d and 4e). The principal components are ordered by how much variability they explain of the first arrivals, and in our case, we only keep the first two components as they explain around 90% of the variability. We did not seek any direct physical interpretation of the principal components, rather used them to describe the variability in the data.

Based on the first two components of the PCA, we apply a novelty detection analysis (Markou & Singh, 2003) using the LocalOutlierFactor class from scikit-learn (Pedregosa et al., 2011). With the novelty detection algorithm, we evaluate whether the observation dataset (in terms of its first two principal components) is an outlier relative to the synthetic forward model response. If the observation dataset is deemed to be a novelty (i.e., outlier), the model can be rejected. Similar approaches to this have been applied in other studies (Hermans et al., 2015; Park et al., 2013; Peeters et al., 2013; Pirot et al., 2019; Scheidt et al., 2015).

3.6. Model Probability Update

The model evidence of the individual process structures $p(Y_{ps} | m_{ps})$ is quantified by the marginal likelihood of the observed data given the model structure and parameters. A more detailed description of the derivation of the likelihood function can be found in section 3.3 and the Supporting Information S2.

The model evidence of the proposed physical structure models $p(Y_{pl} | m_{pl})$ is quantified by a logistic regression classifier using the LogisticRegression class from scikit-learn (Pedregosa et al., 2011). Logistic regression is a supervised machine learning technique that can probabilistically classify data into discrete outcomes.
The classifier is trained on the response from the forward geophysical model (i.e., the model realizations of first arrival data) using a known physical structure and is then applied to the observation data to predict the model evidence of each physical structure.

The motivation of the difference in approach in updating the probability of the process structure and the physical structure is driven by the dimensionality of the testing data. The process structure testing data, i.e., the remote sensing data, are reduced by a knowledge-driven approach to a single metric, that is, the time step when the depression is no longer inundated. This allows direct estimation of the marginal likelihood. The physical structure testing data, i.e., the first arrivals of refraction seismic, have a large dimensionality. Unlike for the process structure, there is no straight-forward knowledge-driven low-dimensional metric, apart from a naïve implementation of an L1 or L2 norm. The logistic regression on a lower dimensional representation of the simulated data avoids defining a summary metric by directly addressing the question: what is the probability that the observation data belong to the ensemble of simulated values.

Figure 4. Workflow for generating one realization of the physical model structure. One of four model structures (Figure 3) is chosen (a), and geometrical parameters are sampled to translate the model structure into a geometrical structure (b). Further, lithological parameters are sampled as input to the rock physics model that translates the lithologies into velocities (c). The velocity structure is used as input to pyGIMLi to obtain a realization of first arrival travel-times (d). The dimensions of the realization of first arrivals are reduced in a principal component analysis (PCA) to two principal components (e). This workflow is repeated 1,000 times for each model structure.
4. Application Results

In this section, we present the results of the applied methodology to the conceptualization of the depressions in Wildman River area. Our goal is to assess the plausibility of different conceptualizations by testing both the process structure and the physical structure with two different, and more importantly, independent types of data. The response of the forward models, a geophysical model and a water balance model, is compared with the observation through the model rejection step and probability update step (Figure 1, and sections 4.1 and 4.2). Last, the posterior probabilities $p(m_c | Y_{ps,pl})$ of the combined process structure and physical structure models are presented, and we evaluate whether the depressions could act as conduits for recharge (section 4.3).

4.1. Process Structure

The water balance forward model (step 3, Figure 1) generated 4,000 synthetic data sets of surface water area response of each of the five depressions (S1 to S5) from which the days the depressions are no longer inundated have been obtained. For each process structure (I to IV), 1,000 responses are available, which represent prior beliefs given the prior parameter values and process structures. The cumulative probability of days, since the end of the wet season, at which the depressions are no longer inundated, is shown in Figure 5 (shaded areas). The $x$ axis is limited to 250 days, as this is the maximum length of the dry season. The cumulative likelihood distributions (defined in section 3.3) for the individual depressions, based on the remote sensing data of open water surface area, are indicated with a solid black line.

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

The results from the four different structures (I to IV) in depressions S1, S2, and S5 look similar, in that their realizations are drying out from around day 100. The realizations pertaining to depressions S3 and S4 dry out at a slower rate, with none of the process structures in S4 drying out before the end of the dry season. This can be attributed to the greater maximum depth (Supporting Information S4) of depressions S3 and S4 and therefore higher starting volumes.

Comparing the model responses (shaded areas in Figure 5) to the likelihood function based on the remote sensing data (solid black line in Figure 5) reveals that the depressions are drying out much faster in reality than most of the forward model responses. Of the simulated responses, process structure II (green area) displays the fastest drying out curve for all five depressions, because of the groundwater recharge component. Similarly, process structure III consistently maintains water in its depressions for a considerably longer period, because of a groundwater discharge component.

The model evidence of the different process structures (I to IV) for the different depressions (S1 to S5), given the remote sensing data, is further discussed in Table 2. Here we show the model evidence values based on marginal likelihoods of the individual models obtained from the likelihood function shown in Figure 5.
Structure II clearly shows the smallest discrepancy with the observations and is outperforming the remaining process structures in all depressions. Any zero model marginal likelihood value indicates process structures whose realizations are all outside of the data-derived likelihood function and therefore obtain a zero marginal likelihood. Such model structures are rejected (step 4, Figure 1), while the model evidence of the remaining structures are estimated (step 5, Figure 1). In depressions S2 and S4, only process structure II remains plausible after the testing exercise, while all four process structures are still plausible for S1 and S5.

Note. The reported model evidence are marginal likelihoods of 1,000 synthetic forward model runs based on a likelihood function defined based on remote sensing data. Entries in bold indicate that the observation (likelihood function) is outside the prior range obtained from the forward water balance model.

### 4.2. Physical Structure

The geophysical forward model (step 3, Figure 1) generated 4,000 synthetic data sets of seismic responses for each of the five depressions (S1 to S5). For each physical structure (A to D), 1,000 responses were available. The interpretation of these datasets by means of principal component analyses is discussed on the basis the first two principal components (dimensions) shown in Figures 6 and 7. The first two dimensions represent between 87 and 93% of the total variance of the seismic response for each of five depressions.

The relationships between velocity (derived from the rock physics model) and the first principal component (Figure 6a) and between different model structures and the second principal component (Figure 6b) have been plotted for S1. The velocity is seen to decrease with increasing values of the first principal component in Figure 6a. Figure 6b is a bee swarm plot showing the distribution of the values of the second principal component for each physical structure. The more realizations for specific values of the principal component, the wider is the distribution along the x axis. The second principal component has lowest values for physical structure A, intermediate values for C and D, and highest values for structure B. This pattern also holds true for the rest of the depressions (except for S5 where the relation between structure and the second principal component is reversed). Figure 6 illustrates that the first component generally explains the velocity, while the second component explains the effect of different physical structures.

![Figure 6](image_url)

**Figure 6.** (a) Relation between the first principal component (Comp 1) and velocities for S1 and (b) the second principal component (Comp 2) and the applied physical structures for S1. The illustrated velocities are mean velocities in the saturated sand zone. The first principal component generally explains the differences in velocity, while the second component explains the differences in model structure. This means that the alternative structures do influence the forward response, but the effect is secondary to the velocity influence.
Comparing the response of models A to B and models C to D, the addition of a higher porosity zone (B and D) is positively correlated with the second principal component. On the other hand, comparing the response of models A to C and models B to D, the addition of a clay layer in the structural model (A and C) does not change the result significantly. Thus, the travel-time response is more sensitive to the high porosity zone than the clay layer.

Figure 7. First and second components of the PCA of the forward geophysical model response for depressions S1 to S5 (rows) for each physical model structure (A to D). Note the different axes for S1 to S3 and S4 to S5. The colored area indicates the frontier outside, which the addition of a data point would be classified as a novelty. If the observation (red cross) plots outside the shaded area, a novelty detection algorithm classifies the observation as a novelty in relation to the model realizations and the model structure can be falsified.
The five rows of Figure 7 present the result for the five depressions (S1–S5), while the four columns present the effect on the seismic response of applying different physical structures (A to D). The input to the rock physics model for generating the seismic responses is the same for all depressions; therefore, the difference between depressions S1 to S5 is solely due to the different inputs to the structural model. For depressions S4 and S5 more variation between forward model response is evident than for the three other depressions (note the different axes on Figure 7). This is probably due to the uncertainty of the water table that is relatively well known in the depressions S1 to S3 but less well known for depressions S4 and S5 (Supporting Information S4).

In each plot in Figure 7, the observed data are illustrated with a red cross. Observation uncertainty has been considered (but not plotted) and generally leads to a variation of less than ±0.005 for the first two principal components. When the observed data point plots outside the forward model response (the shaded area), the data point cannot be predicted by the prior. A novelty detection analysis is applied to automatically classify the observation data as a novelty or an inlier within the prior range (step 4, Figure 1). Recall that when the observation data are classified as a novelty, the prior used to generate the synthetic data is rejected. The colored cloud indicates the frontier between inliers and outliers derived from the model realizations. The rejected models are assigned a zero model evidence in Table 3. For the accepted models, probabilities are calculated with logistic regression classification (step 5, Figure 1).

The results of logistic regression classification of the observed data trained on the realizations are shown in Table 3. Some correspondence can be observed between the observed data location on Figure 7 (red cross) and the performance of the different physical structures; that is, when the response from a model structure is densest around the observation, it performs better. Note, however, that while only two dimensions of PCA are shown in Figure 7, the logistic regression is based directly on the simulated values (thus uses the entire data set). Also, differences in the obtained evidence in Table 3 between points that plot similarly in Figure 7 can be explained by the fact that the sum of the evidence is normalized to unity.

For depressions S1 and S2 the model evidences are almost uniformly distributed, indicating that the alternative structures (A and C for depression S1 and B, C, and D for depression S2) do not generate responses different enough to be able to discriminate between them. Depressions S3 and S4 show a slight preference for structure B (0.37 and 0.67 evidence, respectively), while S5 shows a preference for structure C (0.52 evidence) and generally low preference for structures B and D, including the high-porosity zone in the middle of the depression.

Note. The model evidence is calculated based on logistic regression classification of the observed seismic refraction data. Models rejected through novelty detection (bold) are assigned a zero model evidence.

### 4.3. Posterior Probabilities of Conceptual Models

Comparing the model evidence obtained for the process structure (Table 2) and the physical structure (Table 3), it is apparent that the process structure model evidence is more decisive. This demonstrates the importance of the sensitivity of the data type towards the developed alternative model structures and its ability to discriminate between them. Figure 6 illustrates that the physical structure affects the realizations (based on the second principal component), but that it is secondary to the velocity estimates (based on the first component). The velocities are estimated based on many parameters with relatively wide priors as they are based on literature values (Supporting Information S4). In order to resolve and better discriminate between structures, informative priors are needed. On the other hand, while the difference between the process structures is more pronounced, despite the larger uncertainty of the observations, we can better discriminate between the models.

The posterior probabilities of the conceptual models $p(m|Y_{ps,p})$ of the five depressions, shown in Figure 8, are calculated using Equation 1. The different depressions obtain quite different results, although the expectation was that the depression would behave similarly. This illustrates the importance of considering more than one conceptual model and to have a sufficiently large testing data set.
Of the 16 different conceptual models, only two would allow for the depression to act as conduits of recharge, A-II and B-II. These conceptual models have been marked with a dashed box on Figure 8. In depressions S2, S3, and S4, conceptual model B-II is outperforming the rest of the models, indicating consistency with the depressions being conduits of recharge. In depression S1 the observations plot on the edge of the prior range of realizations for all physical structures (Figure 7), though structure A and C are still classified as probable models. This indicates that the “real” physical structure still might be an unknown. For both the physical structure and the process structure, the model evidence is least decisive for depression S5. For this depression, more model testing would be required in order to discriminate between the conceptual models.

5. Discussion

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

We applied a systematic testing workflow consisting of the following:

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

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

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

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

As for the model testing exercises in other studies (Hermans et al., 2015; Scheidt et al., 2015; Schöniger et al., 2015), we have applied our testing approach in a Bayesian context, where prior model structure probabilities are updated depending on model performance. The Bayesian approach provides a formal framework for iteratively incorporating new data and insights (Figure 1). In each model testing step, models that are inconsistent with data can be removed from the ensemble and conceptual surprises are thereby accommodated.

The insight into the system functioning gained from testing alternative conceptual models can be used in future modeling exercises. With more confidence in the conceptual model, we have more confidence in predictions of (future) system behavior, which provides more robust evidence to underpin environmental management decisions. Further, if a calibration step is considered next, we can be more confident that the obtained parameter values are not biased to compensate for a wrong conceptualization. Thereby giving more confidence in predictions extrapolated beyond the calibration base. Finally, the outcome of the testing framework in terms of updated model probabilities can be used in Bayesian model averaging or selection exercises.

Needless to say, any multimodel approach is limited by our inability to define a collectively exhaustive range of models. However, by applying a rejection step in which model structure performance is compared to data rather than having an intercomparison of performances of different model structures, we avoid assuming that the true model is within the ensemble of the tested models. Indeed, following the workflow from Figure 1, all models can still be rejected. However, as probability sums to unity after updating the probability, we are implicitly assuming that the range of models is collectively exhaustive.

Another limitation of the testing approach is that the result is limited by the information content in the testing data (Nearing et al., 2020). When the information content is low, the definition of the prior probability becomes very important for the result of the testing exercise (Rojas et al., 2009). In relation to the information content of the data, the remaining plausible models after a model rejection step are only conditionally accepted because they have not been proven to be inconsistent with data yet (Oreskes et al., 1994).

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

Nevertheless, our approach is generic and can be applied to any hydrogeological conceptual model where conceptually uncertain components exist. The degree to which the methodology can be applied to large-scale problems is limited by the complexity of the forward model. While running forward models is less intensive than model inversion, the number of model realizations needed to obtain a reliable model probability in step 5 of our workflow increases the computational burden. If the approach is applied to a catchment scale model, where the forward model is temporally and spatially more complex, surrogate modeling may be required to obtain the same number of realizations.

Finally, like other methods based on limits of acceptable change such as Choi and Beven (2007), the applied approach also allows us to uncover conceptual surprises (in case one would reject all models in the model rejection step), but it does not tell us how to deal with them. A conceptual surprise prompts the development of new hypotheses based on model behavior. The new models would, at least indirectly, be based on the
6. Conclusion

We proposed a systematic approach to hydrogeological conceptual model testing, which is needed to increase transparency in the groundwater modeling workflow and seek out conceptual surprises. The approach focuses on using independent data to test models, rather than to develop them. Also, we have emphasized that models should be rejected by comparing performance against data rather than comparing model performance against competing models.

We have applied the approach to the Wildman River area, Australia, involving the testing of the connectivity of widespread, enclosed depressions to groundwater. Our suggested approach is generic and can be applied to any hydrogeological conceptual model where conceptually uncertain component(s) exists.

References

Anderson, M. P., & Woessner, W. W. (1992). The role of the postaudit in model validation. *Advances in Water Resources*, 15(3), 167-173. https://doi.org/10.1016/0309-1708(92)90021-5

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

Bredehoef, J. D. (2005). The conceptualization model problem—Surprise. *Hydrogeology Journal*, 13(1), 37–46. https://doi.org/10.1007/s10040-004-0430-5

Bresciani, E., Cranewick, R. H., Banks, E. W., Battile-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. *Hydrology and Earth System Sciences*, 22(2), 1629-1648. https://doi.org/10.5194/hess-22-1629-2018

Brie, A., Pampuri, F., Marsala, A. F., & Meazza, O. (1995). Shear sonic interpretation in gas fields. *Geophysical Prospecting*, 43(2), 373-382. https://doi.org/10.1111/j.1365-2478.1995.tb00844.x

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society, Series B*, 57, 419–466. Retrieved from https://www.jstor.org/stable/2346440, https://doi.org/10.2307/2346440

Choi, H. T., & Beven, K. (2007). Multi-period and multi-criteria model conditioning to reduce prediction uncertainty in an application of TOPMODEL within the GLUE framework. *Journal of Hydrology*, 332(3-4), 316-336. https://doi.org/10.1016/j.jhydrol.2006.07.012

Clark, M. P., Kavetski, D., & Fenicia, F. (2011). Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resources Research*, 47, W09301. https://doi.org/10.1029/2010WR009827

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., et al. (2008). Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models. *Water Resources Research*, 44, W00B02. https://doi.org/10.1029/2007WR006735

Duan, Q., Ajan, N. K., Gao, X., & Sorooshian, S. (2007). Multi-model ensemble hydrologic prediction using Bayesian model averaging. *Advances in Water Resources*, 30(5), 1371–1386. https://doi.org/10.1016/j.advwatres.2006.11.014

Easey, D., Brocklehurst, P., & Carnavas, M. (2016). *Agricultural land suitability series, Report 2: Soil and land suitability assessment for irrigated agriculture in the Wildman River area*. Darwin, NT, Australia: Northern Territory.

Enemark, T., Peeters, L. J. M., Mallants, D., & Batelaan, O. (2019). Hydrogeological conceptual model building and testing: A review. *Journal of Hydrology*, 569, 310–329. https://doi.org/10.1016/j.jhydrol.2018.12.007

Enemark, T., Peeters, L. J. M., Mallants, D., Batelaan, O., & Sambridge, M. (2019). Hydrogeological conceptual model testing through trans-dimensional sampling of a stochastic water balance model. *Watermark*, 11(7), 1463. https://doi.org/10.3390/w11071463

Engelhardt, I., De Aguiaga, J. G., Mikat, H., Schüth, C., & Liedl, R. (2014). Complexity vs. simplicity: groundwater model ranking using information criteria. *Groundwater*, 52(4), 573–583. https://doi.org/10.1111/gwat.12080

Ferré, T. P. A. (2017). Revisiting the relationship between data, models, and decision-making. *Groundwater*, 55(5), 604–614. https://doi.org/10.1111/gwat.12574

Flinchum, B. A., Holbrook, W. S., Grana, D., Panekian, A. D., Carr, B. J., Hayes, J. L., & Jiao, J. (2018). Estimating the water holding capacity of the critical zone using near-surface geophysics. *Hydrological Processes*, 32(22), 3308–3326. https://doi.org/10.1002/hyp.13260

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

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. https://doi.org/10.1111/j.1467-9868.2011.01203.x

Graham, B. (1985). Surface water resources in the northeastern corner of Wildman River Station (Final report for Water Resources Division project number 2026). Darwin, Northern Territory, Australia.

Gupta, H. V., Clark, M. P., Vrugt, J. A., Abramowicz, G., & Ye, M. (2012). Towards a comprehensive assessment of model structural adequacy. *Water Resources Research*, 48, W08301. https://doi.org/10.1029/2012WR011044

Acknowledgments

This research project was funded through CSIRO’s Deep Earth Imaging Future Science Platform. The seismic datasets for this research are available at https://doi.org/10.17632/dx328m8cg2.1. The Planet Labs Education and Research Program are thanked for providing access to PlanetScope imagery. The authors would like to thank Chris Turnadge, Ursula Zaar, and Steve Tickell for their assistance in the field campaign and Pascal Castellazzi for help in downloading the PlanetScope data and offering insights into its interpretation. The authors would also like to thank Russell Crosbie, Tao Cui, and three anonymous reviewers for review and constructive comments on the manuscript.
Rojas, R. M., Kahunde, S., Peeters, L. J. M., Batelaan, O., Feyen, L., & Dassargues, A. (2010). Application of a multimodel approach to account for conceptual model and scenario uncertainties in groundwater modelling. *Journal of Hydrology*, 384(3–4), 416–435. https://doi.org/10.1016/j.jhydrol.2010.09.016

Rosin, P. L. (2001). Unimodal thresholding. *Pattern Recognition*, 34(11), 2083–2096. https://doi.org/10.1016/S0031-3203(00)00136-9

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

Sambridge, M., Bodin, T., Gallagher, K., & Tkalcic, H. (2013). Transdimensional inference in the geosciences. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371, 20110547. https://doi.org/10.1098/rsta.2011.0547

Scheidt, C., Jeong, C., Mukerji, T., & Caers, J. (2015). Probabilistic falsification of prior geologic uncertainty with seismic amplitude data: Application to a turbidite reservoir case. *Geophysics*, 80(5), M89–M12. https://doi.org/10.1190/geo2015-0084.1

Schöniger, A., Ilman, W. A., Wöhling, T., & Nowak, W. (2015). Finding the right balance between groundwater model complexity and experimental effort via Bayesian model selection. *Journal of Hydrology*, 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 account for conceptual model and scenario uncertainties in groundwater modelling. *Hydrogeology Journal*, 22, 123–137. https://doi.org/10.1007/s10040-013-1087-9

Zyvoloski, G., Kwiciklis, E., Eddehbar, 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. *Journal of Contaminant Hydrology*, 62–63, 731–750. https://doi.org/10.1016/S0169-7722(02)00190-0