Robust Data Worth Analysis with Surrogate Models

by Moritz Gosses1 and Thomas Wöhling2

Abstract

Highly detailed physically based groundwater models are often applied to make predictions of system states under unknown forcing. The required analysis of uncertainty is often infeasible due to the high computational demand. We combine two possible solution strategies: (1) the use of faster surrogate models; and (2) a robust data worth analysis combining quick first-order second-moment uncertainty quantification with null-space Monte Carlo techniques to account for parametric uncertainty. A structurally and parametrically simplified model and a proper orthogonal decomposition (POD) surrogate are investigated. Data worth estimations by both surrogates are compared against estimates by a complex MODFLOW benchmark model of an aquifer in New Zealand. Data worth is defined as the change in post-calibration predictive uncertainty of groundwater head, river-groundwater exchange flux, and drain flux data, compared to the calibrated model. It incorporates existing observations, potential new measurements of system states ("additional" data) as well as knowledge of model parameters ("parametric" data). The data worth analysis is extended to account for non-uniqueness of model parameters by null-space Monte Carlo sampling. Data worth estimates of the surrogates and the benchmark suggest good agreement for both surrogates in estimating worth of existing data. The structural simplification surrogate only partially reproduces the worth of "additional" data and is unable to estimate "parametric" data, while the POD model is in agreement with the complex benchmark for both "additional" and "parametric" data. The variance of the POD data worth estimates suggests the need to account for parameter non-uniqueness, like presented here, for robust results.

Introduction

Management of groundwater systems is mandatory for safe and sustainable usage of the natural resources. Aiding the management of aquifers around the world are models of the hydrogeological systems (e.g., Cirpka et al. 2004; Taormina et al. 2012; Roy et al. 2016; Wöhling et al. 2018). Calibration of model outputs against measured system data is often undertaken to estimate sensible model parameters for reliable predictive performance of the model (e.g., Beven and Binley 1992; Moore and Doherty 2005). While this process would optimally include many diverse measurements, data about the system of interest is often both scarce and uncertain due to time and cost constraints. Thus, maximizing informational gain by targeted monitoring design is an important task in efficient model building and application (cf Fienen et al. 2010).

Many researchers in hydrological sciences have taken on the question of worth of data in recent years (e.g., Cirpka et al. 2004; Fu and Gómez-Hernández 2009; Dausman et al. 2010; Wöhling et al. 2016; Siade et al. 2017; Sreekanth et al. 2017; Knowling et al. 2020). Tangentially, data worth estimation is often linked to the topic of monitoring network design, which includes the costs of potential measurements as well (e.g., Erickson et al. 2002; Feyen and Gorelick 2005). In this study, we will neglect the aspect of cost and focus solely on the informational gain through data, though.

Worth of data has been defined as reducing/increasing the uncertainty associated with model predictions (e.g., Beven and Binley 1992; Cirpka et al. 2004; Dausman et al. 2010) or as influencing the determination of model parameters (e.g., Siade et al. 2017). Another distinction is the type of data that is considered. Fienen et al. (2010) differentiate between three different types, which all have been investigated in hydrological research. First is the analysis of existing data, where the increase in uncertainty through the exclusion of existing measurements is
estimated (e.g., Dausman et al. 2010; Fienen et al. 2010). This allows an inspection of existing monitoring networks for redundant information. Second, and most common, is the analysis of worth of “additional” data (e.g., Fienen et al. 2010; Wöhling et al. 2016), with “additional” referring to data not part of the current measurement network as defined by the calibration data set. This aims to answer the common question of which new data to collect to improve the accuracy of model predictions. Third, the potential knowledge of certain model parameters and its effect on reducing predictive uncertainty can be analyzed (e.g., Feyen and Gorelick 2005; Moore and Doherty 2006; Fienen et al. 2010). This “parametric” data can be used to identify parameters highly sensitive to predictions of interest. This is of especially high interest where the physical properties represented by the parameter in questions could potentially be measured. Wöhling et al. (2016) even extended the data worth methods to simultaneous analysis of multiple data points and types.

In analyzing worth of data for predictive uncertainty estimation, researchers have applied many different methods for the underlying predictive uncertainty estimation, both first-order second-moment methods (e.g., Cirpka et al. 2004; Dausman et al. 2010; Fienen et al. 2010) and stochastic methods (e.g., Feyen and Gorelick 2005; Fu and Gómez-Hernández 2009). Stochastic analysis is often unfeasible due to the high number of model runs required, while first-order second-moment based analysis is subject to (unknown) error. A number of authors have tried to tackle those problems by using methods that employ targeted sampling of the parameter space via null-space Monte Carlo, for example (Siade et al. 2017).

Another recent method to further the applicability of data worth analysis has been the combination with surrogate models. Surrogate models is a term that encompasses a multitude of simplification-methods and stand-ins, with the common goal to “increase computational efficiency” (Asher et al. 2015). Usually, they are divided in three distinct categories (cf Asher et al. 2015): projection-based methods (e.g., Vermeulen et al. 2004a; McPhee and Yeh 2008; Pasetto et al. 2013; Boyce and Yeh 2014; Ushijima and Yeh 2015; Stanko et al. 2016), data-driven methods (e.g., Khu and Werner 2003; Yoon et al. 2011; Taormina et al. 2012; Roy et al. 2016; Fan et al. 2020; Yin and Tsai 2020), and structural simplification methods (e.g., Sun 2008; Doherty and Christensen 2011; Watson et al. 2013; von Gunten et al. 2014; Knowling et al. 2020).

The run time reduction through surrogate models has been used to accelerate different modeling tasks such as forecasting (e.g., Daliakopoulos et al. 2005; Adamowski and Chan 2011), inverse modeling (e.g., Winton et al. 2011; Boyce and Yeh 2014; Burrows and Doherty 2016), stochastic simulation (e.g., Pasetto et al. 2013), uncertainty analysis (e.g., Doherty and Christensen 2011; Burrows and Doherty 2016), and optimization of management strategies (e.g., McPhee and Yeh 2008; Mulligan and Ahlfeld 2016) or experimental design (e.g., Ushijima and Yeh 2015). Besides reducing model run times, surrogate models can also possibly reduce numerical instability (Doherty and Christensen 2011), aid in the analysis of (in-) sensitive parameterization (Watson et al. 2013) or can be used for the determination of adequate model structures (Matott et al. 2009).

In the field of data worth analysis for hydro systems, surrogate models have been applied in recent years for optimization of remediation strategies (e.g., Sbai 2019; Yin and Tsai 2020), optimal design of measurement networks (e.g., Sreekanth et al. 2017; Sreekanth et al. 2020), or general groundwater management/exploitation optimization (e.g., Roy et al. 2016; Fan et al. 2020). Knowling et al. (2020) used structurally simplified surrogates in combination with a full benchmark model to estimate data worth and the effects of data on prediction confidence.

To our knowledge, surrogate models have not been applied to data worth analysis methods that estimate the effects of parameter non-uniqueness in a robust way, but are fast enough to allow for thorough benchmarking regarding different model predictions and data types. We, therefore, combine two different methods to alleviate the high run times associated with robust data worth analysis in this study:

- Combination of fast first-order second-moment uncertainty quantification and data worth estimation following Doherty (2016) with null-space Monte Carlo methods (Doherty 2016) for efficient sampling of the calibrated parameter space to account for parameter non-uniqueness, and
- Application of two different surrogate models for data worth analysis, with a thorough comparison to results of a complex benchmark model to determine their applicability for different predictions and data types. For this, we chose a model that, while physically based and highly parameterized, is still fast enough to allow the application of the robust data worth method presented here to provide a benchmark for comparison.

Furthermore, we link the results to the simplification error analysis as presented in Gosses and Wöhling (2019) and Gosses (2020) to address possible connections between predictive performance and suitability for data worth analysis for both surrogate models.

This paper is structured as follows: in section Methods, we combine the methodology of the first-order second-moment uncertainty analysis and parameter set generation with null-space Monte Carlo to a robust data worth estimation. We also present the case study along with its complex model and both surrogates. Furthermore, this section contains a short summary of the method of the simplification error analysis. Section Results summarizes the simplification errors of the surrogate predictions and holds the results of the application and discussion of the data worth analysis for both surrogates and the benchmark model regarding existing, “additional” and “parametric” data. Section Discussion holds the interpretation of the results, while we summarize the main findings and draw conclusions in section Summary and conclusions.
Methods

Robust Data Worth Analysis

In this section, we present the robust data worth analysis applied in this work. We propose a combination of first-order second-moment uncertainty and data worth estimation methods with a highly efficient method to generate calibrated model parameter sets to a fast, but robust data worth analysis method in light of parameter non-uniqueness. First, we present the theory behind the first-order second-moment predictive uncertainty estimation methods following previous work by Moore and Doherty (2006); Dausman et al. (2010); Doherty (2016); Wohling et al. (2016), among others. Then, we present how to apply these methods to compute the worth of three different categories of data (existing, “additional” and “parametric”) in regards to their effect on predictive postcalibration uncertainty. We next summarize the methodology for efficient generation of calibrated parameter sets via null-space perturbation using tools developed by Doherty (2016). Finally, we combine the first-order second-moment uncertainty and data worth estimation methods with the efficient parameter set generation into the robust data worth analysis applied in this work.

First-Order Second-Moment Uncertainty Estimation

The presented method for the first-order second-moment estimation of uncertainty in regards to model predictions follows the strategy developed, extended and applied by Moore and Doherty (2006); Dausman et al. (2010); Doherty (2016); Wohling et al. (2016) and others. Thus, we only provide an abbreviated overview of the major definitions and equations to follow its application here. More exhaustive derivations and discussion can be found in the original literature.

We define a linearized model:

$$h = Zk + \varepsilon,$$  \(\text{(1)}\)

where \(h\) denotes the calibration data set, \(Z\) is a matrix representing the model’s action under calibration, the vector \(k\) holds the model parameters and \(\varepsilon\) is a vector of (random) measurement noise. In reality, groundwater models are usually nonlinear, which effects the results, but does not invalidate them.

A prediction made by this model is defined as:

$$s = y^T k,$$  \(\text{(2)}\)

where scalar \(s\) is the prediction and the vector \(y\) holds the sensitivities of the prediction to the model parameters \(k\). From the definition of a model prediction as given in Equation (2), one can describe the uncertainty of the prediction by its variance \(\sigma_s^2\) given as:

$$\sigma_s^2 = y^T C(k)y,$$  \(\text{(3)}\)

where \(C(k)\) is the precalibration covariance matrix of the parameters. Note that this is an estimation of the precalibration uncertainty, as it is not subjected to data. To compute the postcalibration uncertainty from this term, we start by combining Equation (2) (the definition of a model prediction) and Equation (1) (the definition of the linear model) to get:

$$\begin{bmatrix} s \\ h \end{bmatrix} = \begin{bmatrix} y^T & 0 \\ Z & I \end{bmatrix} \begin{bmatrix} k \\ \varepsilon \end{bmatrix},$$  \(\text{(4)}\)

where \(I\) is the identity matrix. Assuming multivariate normal distribution of the model parameters \(k\) and the error \(\varepsilon\), the covariance of Equation (4) is thus defined as:

$$C(\begin{bmatrix} s \\ h \end{bmatrix}) = \begin{bmatrix} y^T C(k)y & y^T C(k)Z^T \\ ZC(k)y & ZC(k)Z^T + C(\varepsilon) \end{bmatrix}.$$  \(\text{(5)}\)

Using the definition of the conditional variance for Equation (5), which gives the computation of \(C_s \equiv \sigma_s^2\), the variance of the prediction \(s\) conditional of the calibration data \(h\) is:

$$\sigma_s^2 = y^T C(k)y - y^T C(k)Z^T$$

$$\times (ZC(k)Z^T + C(\varepsilon))^{-1}ZC(k)y.$$  \(\text{(6)}\)

Note how this definition of the predictive postcalibration uncertainty is computed as a reduction from the precalibration uncertainty \(y^T C(k)y\), as defined in Equation (3) through the calibration with the data \(h\).

Computing \(\sigma_s^2\) not only necessitates \(Z\) (the Jacobian matrix of the calibrated model, which can easily be calculated) and \(y\) (the vector of the sensitivities of \(k\) to the prediction of interest \(s\)), but also the covariance matrices of the parameters \(C(k)\) and observations \(C(\varepsilon)\), respectively. Where spatially distributed, correlated parameters (like pilot points of hydraulic conductivity) are used, PEST (Doherty 2016) provides tools for the computation of the accompanying parts of the \(C(k)\) matrix. For non-correlated parameters, we fill the diagonal of \(C(k)\) with estimations of the variance of the respective parameter. While observations are usually thought to be correlated, the nature of this correlation is usually unknown. Due to this, and the simplicity of it, we assume observations to be uncorrelated, so \(C(\varepsilon)\) is defined as a diagonal matrix proportional to the estimated measurement errors of the observations. Note that the model can include a multitude of “observations” that do not have actual measurements associated to them (model predictions or similar) or are not used in the calibration process. These have weights of zero assigned to them. Throughout this work, predictive uncertainty refers to the predictive postcalibration uncertainty as defined by Equation (6). Predictive uncertainty calculated while including the whole calibration data set is used as the base case for the estimation of data worth in the following.

Worth of Data

As presented earlier, we follow Fienen et al. (2010) by differentiating three types of data that is of interest...
in the analysis of data worth in groundwater modeling. All three types can be incorporated by the first-order second-moment uncertainty estimation as presented in Equation (6) through the following:

1. Existing data:
In the context of this work, all data that was used as part of the calibration data set of the model under consideration is defined as existing data. This is regardless of the type of measurement, but does not include data used for model verification or data used for regularization of model parameters. The worth of existing data can be computed via Equation (6) by setting the weight $w$ to compute the diagonal entry of $C(\varepsilon)$ of the observation of choice to zero and comparing the resulting $\sigma_{sh}^2$ to the base case. The potential increase in predictive uncertainty is the measure of data worth in this case.

2. “Additional” data:
“Additional” data is defined as measurements of system states that were not part of the calibration data set. This can include both data reserved for model verification as well as non-existing measurements, as the value of the measurement is not necessary for the evaluation of its worth. This is undertaken by adding the measurement type, location and time of choice to the observation data set and subsequent recalculation of the Jacobian matrix (to gain $Z$ and $y$ with entries for this “additional” data). Then, the “additional” data can be added to the uncertainty estimation process by assigning them diagonal entries in proportionality to their estimated measurement error in the $C(\varepsilon)$ matrix. It’s worth is the (potential) reduction in predictive uncertainty through the inclusion of this “additional” data in comparison to the base case. Due to the independence of an actual measurement of the system state in question, the method can be used to pre-emptively estimate the worth of yet unknown data in reducing the postcalibration uncertainty of model predictions.

3. “Parametric” data:
The term “parametric” data incorporates the assumed knowledge of system properties defined as model parameters, not as (potential) observations. By setting the variance of the parameter in question to zero in $C(k)$, the model assumes perfect knowledge of this parameter. The resulting (potential) reduction in $\sigma_{sh}^2$ can again be compared to the base case as an estimate of the worth of this “parametric” data for postcalibration uncertainty of the chosen model predictions. Note that while such perfect parameter knowledge cannot be obtained in reality, it still gives significant insight into the relative importance of model parameters for the uncertainty of the prediction of interest. Where model parameters can potentially be measured, the variance of such parameters can be set to the measurement error instead of zero.

Generating Calibrated Parameter Sets — Null-Space Parameter Perturbation

The data worth estimation methods presented above are linear in respect to the calibrated parameter set, which requires the model to be calibrated, with $k$ representing the calibrated parameter set. Potential parameter non-uniqueness is accounted for by null-space Monte Carlo sampling techniques as implemented in the PEST (Doherty 2016) toolbox.

It consists of the following steps:

1. Creation of $i$ random parameter sets $k_i$ based on the prior parameter probability distribution as defined by $C(k)$.
2. The original calibrated parameter set $k$ is subtracted from the random parameter sets $k_i$, yielding $i$ parameter difference sets $k_i - k$.
3. Projection of the parameter difference sets $k_i - k$ onto the null-space of $k$ to generate $(k_i - k)^{ns}$.

For a linear model, this process would result in different parameter sets that all calibrate the model at the same precision as the original calibrated parameter set $k$, as they only differ from $k$ through components in its null-space $(k_i - k)^{ns}$, which, by definition, do not affect the calibration data set (and thus the precision of calibration). Due to model non-linearity, solution space and null-space cannot be separated completely, resulting in some solution space component in $(k_i - k)^{ns}$, which alters the model-measurement fit of the parameter sets created in step 4 (usually to a less-calibrated fit). To alleviate for this, a final step is added to the process:

4. Addition of the null-space projected parameter difference sets $(k_i - k)^{ns}$ onto the original calibrated parameter set, $k + (k_i - k)^{ns}$.

Again, for a linear model, this re-calibration step would only affect the solution space of the model, thus keeping the null-space deviation created earlier fully intact. Dealing with model non-linearity, this cannot be assured. Thus, while the re-calibration step is necessary to ensure that the random parameter sets generated calibrate the model sufficiently, this step can also affect the null-space deviation of the newly generated parameter sets. Too rigid a re-calibration could thus potentially falsely narrow the variability of the random parameter sets. Therefore, the recalibration has to be done cautiously, trying to generate sufficiently calibrating parameter sets while not narrowing the variability of the total of all sets too much. This process can therefore not capture the total variability of potential calibrating parameter sets due to model non-linearity, but only approach it. Nonetheless, the method is far less time-consuming that generating calibrated parameter sets from differing starting points by
conventional calibration, especially in conjunction with surrogate models due to their low run times.

Another possibility, besides re-calibration, to address the influence of the solution space on the perturbed parameter sets is discarding all parameter sets \( k + (k_i - k)^{\text{ps}} \) that do not calibrate the model satisfactorily, measured by a threshold objective function of choice, for example. Furthermore, generation of parameter sets \( k + (k_i - k)^{\text{ps}} \rightarrow \text{rc} \) can often be computationally less demanding than re-calibration to generate parameter sets \( k + (k_i - k)^{\text{ps}} \rightarrow \text{rc} \) with sufficient model-to-measurement fit. As above, the downside of narrowing the total variability of potential calibrating parameter sets remains. This second option was chosen for this study, and is presented in more detail in the section “setup of the test case” below.

Robust Data Worth Analysis

For the application of a robust data worth analysis method in this study, we combined the first-order second-moment uncertainty and data worth estimation with the multiple calibrating parameter sets generated as presented above. First-order second-moment data worth estimation and efficient calibrating parameter set generation are combined through the following steps:

1. Generation of a (base) parameter set \( k \) by calibration of the model.
2. Generation of \( i \) random null-space perturbed parameter sets \( k + (k_i - k)^{\text{ps}} \rightarrow \text{rc} \), which calibrate the model to a sufficient degree, via the method presented directly above.
3. Recalculation of \( Z \) and \( y \) for each of the \( i \) perturbed parameter sets \( k + (k_i - k)^{\text{ps}} \rightarrow \text{rc} \).
4. Calculation of the basic predictive uncertainty \( \sigma_{ij}^2(i) \) for \( j \) predictions of interest \( s_j \) for each of the \( i \) parameter sets \( k + (k_i - k)^{\text{ps}} \rightarrow \text{rc} \).
5. Calculation of the worth of existing, “additional” and “parametric” data as increase or decrease of the basic predictive uncertainties as explained above. This can be done for all (combinations of) parameters and measurements of interest.
6. Aggregation of the \( i \) different data worth results for each prediction of interest (and each data type) by calculation of mean and 90% quantiles of data worth. Other statistical measures could obviously be applied here, as well.

The resulting data worth analysis, while based on first-order second-moment methods, creates more robust data worth estimations due to an approximation of the effects of parametric uncertainty while remaining reasonably time-consuming.

Case Study: The Wairau Plain Aquifer Model

This subsection gives a short overview of the case study utilized in this work, along with the complex model used as a benchmark for the performance assessment of the surrogate models. Both the Wairau Plain Aquifer and the complex model are described in more detail in Wöhling et al. (2018). For detailed information on the Wairau Aquifer, its location, general settings, and characteristics, please refer to Wöhling et al. (2020).

The Wairau Plain Aquifer

The data worth analysis presented in this paper was applied to an existing research area: the Wairau Plain Aquifer in New Zealand. Located in the northern part of the South Island near the city of Blenheim, the unconfined Wairau Plain Aquifer is mainly recharged by an approximately 27 km long section of the Wairau River and is between 20–35 m thick. A confining, wedge-shaped layer of marine sediments lies on top of the aquifer in the eastern part of the Wairau Plain, which pushes the groundwater toward the surface through several springs and back into the north-eastern part of the Wairau River.

The monitoring network consists of four flow gauging sites in the Wairau River, one flow (stage) gauge of Spring Creek, and several groundwater observation bores on the Wairau Plain. Climate data (precipitation, evapotranspiration) is sourced from a virtual climate station network operated by NIWA (Tait et al. 2006). Groundwater abstraction and groundwater recharge on the Wairau Plain are estimated by a soil-water balance model (Wöhling et al. 2018).

Complex MODFLOW Model of the Wairau Plain Aquifer (CM)

The assessment of surrogate models in data worth analysis necessitates a complex benchmark model (labeled CM) as a point of comparison. This model has to be simple enough to allow application of the robust data worth methods presented here in reasonable time frames to function as the benchmark. The complex model of the Wairau Plain aquifer that we use as the benchmark is a transient coupled surface water-groundwater model realized in MODFLOW-NWT (Niswonger et al. 2011) using the ModelMuse (Winston 2009) GUI. It consists of 6360 active model cells, spanning the approximately 85 km² model area with a 200 × 200 m horizontal grid and three vertical layers representing the major geological units, as shown in Figure 1a. Data from July 1, 2013 to November 6, 2016 were used to spin up, calibrate and validate the model, resulting in a total of 1225 daily time steps. The eastern, downstream model boundary was set as a constant groundwater flux boundary simulating the groundwater flux into the sea. The northern boundary, set by the Wairau River, and the southern boundary, a no-flow boundary set perpendicular to estimations of the groundwater flow field, meet in the western, upstream end of the model domain. The Wairau River was implemented via the stream-flow routing package (Niswonger and Prudic 2005), the springs in the eastern model area were simulated as drains, and the groundwater recharge and abstraction estimates of the soil-water balance model were used as model forcings over the domain.

Ninety pilot point parameters each were used to discretize both hydraulic conductivity \( K_H \) and specific yield \( S_y \) of the three geological units for a total of 180 parameter values. The interpolation of these values
followed an exponential variogram derived from pumping tests, and Tikhonov regularization as implemented in PEST (Doherty 2016) was used to ensure consistency with this variogram for the pilot point values. Furthermore, for each of the three geological units, a single specific storage parameter $S_s$ and anisotropy ratio $f_a$ was defined. For the Dillon’s Point aquitard, a single parameter was used for all four geological properties. Finally, bed conductivities of the Wairau River ($K_R$) and the different springs ($K_D$) were defined as parameters, for a total of 207 model parameters of the full MODFLOW model, as summarized in Table 1. Model predictions range from groundwater heads at other measurement points to piece-wise river-groundwater exchange fluxes and spring discharges.

**Surrogate Models**

This subsection gives a short overview of the two surrogate models applied for data worth analysis. Surrogate 1, the simplified MODFLOW model, is presented in more detail in Gosses and Wöhlung (2019), while the theory underlying surrogate 2, the linearized POD model, can be found in Vermeulen et al. (2004) or Gosses et al. (2018), for example.

**Surrogate 1: Simplified MODFLOW Model SM$_{1,SMm}$**

The first of two different surrogate models belongs to the category of structural simplification methods and is a simplified version of the CM model also realized in MODFLOW, called SM$_{1,SMm}$. Most of the model specifications, boundaries and data are identical to the CM model, with the following differences:

- Reduction of the number of active cells from 6360 to 756 by universally coarsening of the horizontal grid from $200 \times 200$ m to $600 \times 600$ m.
- Replacement of the spatially distributed $K_H$ and $S_y$ parameter fields by single parameter values per geological unit.
- Aggregation of the river bed sections, reducing the associated bed parameters from 12 to 3 and the total parameter number from 207 to 24.

![Figure 1. (a) Complex CM and (b) simple SM$_{1,SMm}$ MODFLOW models with the main geological features.](Image)
CM model. A snapshot set was created by running a full estimations with satisfactory accuracy (not shown).

reproduces simulated model outputs and uncertainty that this linearized representation of the CM model

Using the groundwater heads of

• Using the river head values simulated by the SFR

• Approximating the (originally unconfined) groundwater

• Using the groundwater heads of t-1 for determining connected drain cells.

Some preliminary studies were conducted to ensure that this linearized representation of the CM model reproduces simulated model outputs and uncertainty estimations with satisfactory accuracy (not shown).

Basic POD reduction was applied to this linearized CM model. A snapshot set was created by running a full forward simulation and taking the simulated groundwater heads at every 50 days \( \mathbf{H}_t = [d_{t,11}; d_{t,55}; d_{t,101}; \ldots; d_{t,1201}; d_{t,1225}] \), and subsequently singular value decomposition (SVD) was used to generate both singular values and left and right singular vectors. Following the methodologies common in literature, we chose a relative retained variance of 99.9% to retain seven left singular vectors for the subspace projection, and thus as the surrogate model dimension. The projection matrix generated from the original calibrated CM model parameter set was used throughout this study—no regeneration of \( \mathbf{P} \) was undertaken for the different parameter sets used for data worth analysis.

Overall Model Setup and Run Time Comparison

The CM benchmark and both surrogate models were set up very similar to allow comparison for their predictive uncertainty and data worth estimations in this study. This setup, along with a comparison in computational run times, is presented in this section.

All three models were calibrated using the Gauß-Marquardt–Levenberg method with the parameter estimation software PEST (Doherty 2016), as described for the CM benchmark in Wöhlung et al. (2018) in detail. Generally, the same procedure was followed for all three models here. Five continuous groundwater head measurements (labeled 1-5) were used for parameter calibration, along with measured discharge at the Spring Creek (SC) spring. Differential Wairau River flow gaugings and estimates of mean discharge for the other springs were used as “soft” targets during the model calibration. Calibration adjusted the full 207 model parameters for the CM benchmark and the SM2,POD surrogate, while only 24 model parameters were estimated for the SM1,sMm surrogate.

A major factor of surrogate modeling, as presented earlier, is the run time reduction. As can be seen in Table 2, the SM1,sMm surrogate runs about five times faster than the CM benchmark for a single forward run, while calibration of the SM1,sMm surrogate takes about a third of

### Table 1

Parameterization of the Wairau MODFLOW models (CM and SM1, sMm)

| Parameter                                      | Type [unit] | Property | Range       | CM model                          | Type | Quantity | SM1,sMm model | Type | Quantity |
|------------------------------------------------|-------------|----------|-------------|-----------------------------------|------|----------|---------------|------|----------|
| Hydraulic conductivity \((K_H)\) [m/d]          | Layer 1     |          | 1–3500      | Pilot Points                      |      | 26       | Zones         |      | 1        |
|                                                | Layer 2     |          | 1–2500      | Pilot Points                      |      | 31       | Zones         |      | 1        |
|                                                | Layer 3     |          | 1–3500      | Pilot Points                      |      | 33       | Zones         |      | 1        |
|                                                | Aquitard    |          | 0.1–50      | Zones                             |      | 1        | Zones         |      | 1        |
| Specific yield \((S_y)\) [-]                  | Layer 1     |          | 1e⁻³–0.35   | Pilot Points                      |      | 26       | Zones         |      | 1        |
|                                                | Layer 2     |          | 1e⁻³–0.35   | Pilot Points                      |      | 31       | Zones         |      | 1        |
|                                                | Layer 3     |          | 1e⁻³–0.35   | Pilot Points                      |      | 33       | Zones         |      | 1        |
|                                                | Aquitard    |          | 1e⁻⁶–1e⁻³   | Zones                             |      | 1        | Zones         |      | 1        |
| Specific storage \((S_s)\) [m⁻¹]               | Each        |          | 1e⁻⁷–1e⁻³   | Zones                             |      | 4        | Zones         |      | 4        |
| Anisotropy ratio \((f_a)\) [-]                 | Each        |          | 1–10        | Zones                             |      | 4        | Zones         |      | 4        |
| River bed conductivity \((K_R)\) [m/d]         |             |          | 1e⁻³–0.2    | Sections                          |      | 12       | Sections       |      | 3        |
| Drain bed conductivity \((K_D)\) [m/d]         |             |          | 1e⁻⁴–1000   | Sections                          |      | 5        | Sections       |      | 5        |
| Total                                          |             |          |             |                                    |      | 207      |               |      | 24       |
the time of the CM benchmark. The SM$_{2,POD}$ surrogate reduces model run times by a factor of 25 for a single forward run and by a factor of 12 for the calibration process. Note that this includes the single CM model run necessary to compute the projection matrix for the generation of the SM$_{2,POD}$ surrogate. The superior run time of a single SM$_{2,POD}$ surrogate model run keeps the overall time for calibration smaller than the time of the SM$_{1,sMm}$ surrogate, despite its smaller number of model parameters.

Looking just at the time to calculate the Jacobian matrix of the models (step 3 in Table 2), the effect of the smaller number of model parameters of the SM$_{1,sMm}$ surrogate is visible, as the calculation of the Jacobian matrix takes a single model run per model parameter. The SM$_{1,sMm}$ surrogate is 45 times faster than the CM benchmark in calculating $Z$ and $y$, while the SM$_{2,POD}$ surrogate is about 27 times faster.

Finally, Table 2 lists the total run times of the three models to perform the robust data worth analysis as presented in this study. With a total run time of about 19 days, robust data worth analysis with the CM model was still possible, which was necessary for its utilization as a benchmark for the two surrogates. As the total run time for the data worth analysis depends mainly on the time to recalculate the Jacobian matrix for each parameterization, the SM$_{1,sMm}$ surrogate is the faster than the SM$_{2,POD}$ surrogate, reducing the run time by factor 45 instead of 27. Generally, both surrogates are able to perform the robust data worth analysis in less than a day, while the CM benchmark can perform the robust data worth analysis in a time frame of less than 20 days to provide the necessary results for comparison for the two surrogates.

### Setup of the Test Case

This section provides a short delineation of the general setup used for applying the robust data worth analysis presented above to the case study and its models.

Due to the prediction-specific nature of the data worth estimation, diverse model predictions were chosen as targets. All predictions (except prediction $S_{917}$, the total groundwater storage) are displayed in Figure 1:

(a) $H_{WR}$: the minimum groundwater head during the model time in the (measured) Wratts Rd. well.

(b) $H_{N15}$: the minimum groundwater head in a potential new well used as “additional” data.

(c) $S_{917}$: total groundwater storage in the model domain at time $t = 917$ days, taken from the overall mass balance.

(d) $Q_{ex}$: total mean river-groundwater exchange flux between gauging stations at Rock Ferry and at SH1.

(e) $Q_{ex3}$: mean river-groundwater exchange flux between the last two gauging stations at Wratts Rd. and at SH1.

(f) $Q_{sp3}$: mean groundwater flux into the Southern Streams, the combined springs and Opaoa River in the southern part in the model.

Estimation of the existing groundwater measurement network is undertaken by sequentially excluding one groundwater well from the calibration data set. While the analysis of “additional” data worth allows the addition of new potential measurements of any type and location, we focused solely on groundwater head measurements at potential new wells. These potential measurements were evaluated on a raster (locations are marked as dots in Figure 4) and assumed to be measured continuously over the calibration time period of the model, in accordance with the existing real groundwater head measurements. Each “additional” groundwater well is added to the full existing calibration data set one at a time with the same weight as assigned to the existing, real groundwater head measurements (see Wöhling et al. 2018). Similarly, analysis of “parametric” data is undertaken by sequentially assuming perfect knowledge of a single parameter in combination with the full existing calibration data set. Please note that this data worth methodology could be extended to multiple simultaneous new measurements of different type and location as described in Wöhling et al. (2016).

In regard to the multiple random calibrated parameter sets providing robustness, $i = 1000$ parameter sets $k_i = (k_i - k)^{jth}$ were generated for each of the three models. From these $i = 1000$ parameter sets each, $\tilde{i} = 100$ sets where chosen as suitable for the final application. They were deemed to sufficiently calibrate the model with objective function values of less than twice the objective function value of the original calibrated parameter set.

### Simplification Error Analysis

This subsection addresses the simplification error that arises from the use of surrogate models. Here, we define

| Approximate run time for | CM model | SM$_{1,sMm}$ model | SM$_{2,POD}$ model |
|--------------------------|----------|---------------------|--------------------|
| 1. A single forward run   | 75 s     | 15 s                | 3 s                |
| 2. Model calibration      | 36 h     | 12 h                | 3 h$^1$            |
| 3. Generation of $Z$ and $y$ | 4.5 h | 6 min               | 10 min             |
| 4. Generation of calibrated parameter sets | 10 s | 10 s                | 10 s               |
| 5. Total robust data worth analysis | 19 d | 10 h                | 17 h               |

$^1$Includes a single CM forward run for the creation of the projection matrix.
the simplification error of a surrogate as the prediction-specific error between complex model and surrogate model. As this paper presents analysis of data worth with surrogate models for specific model predictions, it is important to estimate the general trustworthiness of such surrogate model predictions. This might influence the precision of data worth estimates of the surrogates regarding those predictions.

We analyze the simplification error of surrogate model predictions in accordance to our previous work (Gosses and Wöhling 2019; Gosses 2020). Therefore, only a short summary of the method is presented here. First, the complex and surrogate model are linked by calibration of the surrogate against complex model outputs for a variety of semi-random complex model parameter fields. Calculation of the prediction pairs of both complex and surrogate model for each of these parameter realizations can be displayed as scatter plots, from which the variance of surrogate model predictions can be estimated. This variance defines the simplification error of said surrogate model prediction. Extensive discussion can be found in Doherty and Christensen (2011) and Watson et al. (2013), which originally presented the method applied here. Furthermore, we extended the method in Gosses and Wöhling (2019) and applied it to two different surrogate models, one of them the SM1,sMm surrogate used in this work. We undertook the same analysis for the SM2,POD surrogate (Gosses 2020) and will briefly summarize the relevant results later.

Generally speaking, small simplification errors indicate good reproduction of the complex model predictions by the surrogate, while large simplification errors reveal shortcomings of the surrogate to estimate the prediction of interest with confidence (again, in comparison to the complex benchmark). Furthermore, the analysis allows to estimate a mean bias of the surrogate prediction, another indication for structural deficiencies of the surrogate.

**Results**

We start the section with a short summary of the simplification error results for both surrogate models and how they relate to the data worth analysis. Then, the results of the robust first-order second-moment data worth analysis are presented subsequently in three parts, following the three types of data under consideration: worth of existing, “additional” and “parametric” data.

**Prediction-Specific Simplification Errors**

Figure 2 presents the scatter plots of the simplification error analysis for both surrogate models for four of the six model predictions under consideration in this study. The structural simplification surrogate SM1,sMm shows relatively small simplification errors for three of the four predictions. Only for the prediction of the flux of the low-land springs called southern streams, Qsp3, do the results suggest a high simplification error and bias of the surrogate SM1,sMm. Thus, it seems to be able to reproduce most predictions of the complex benchmark model with little error along a wide range of possible parameterizations.

The POD surrogate SM2,POD, only exhibits small simplification error for the prediction of overall Wairau River-groundwater exchange flux, Qex. For the other three predictions, the simplification error of the POD surrogate is large, and the prediction of Qsp3 especially is extremely biased. This does not give confidence in the POD model regarding the reproduction of these
predictions. Its inability to predict negative exchange fluxes between Wratts Rd. and SH1 gauging stations, $Q_{exc3}$, and the extreme bias of the $Q_{sp3}$ suggest that the POD surrogate cannot adequately represent the interactions between surface water and groundwater in the eastern part of the model, especially when considering multiple parameterizations.

From these results, one could assume that data worth estimates will be more similar to the benchmark when made by surrogates with small simplification error, and less similar when the surrogate exhibits big simplification error. This assumption will be tested in the following.

**Worth of Existing Data**

We first present the results of the data worth analysis for the existing groundwater head measurement data used in the calibration data set of both surrogate models (SM$_{1,sMm}$ and SM$_{2,POD}$) and the complex model (CM). Figure 3 shows the data worth for all three models for two different predictions, $S_{917}$ (total aquifer storage) and $H_{WR}$ (minimum head in a groundwater well) for the five groundwater wells used in the calibration data set. The box plots map data worth as mean and variance of the calculations with the 100 different parameterizations. Note that no absolute values are given, as these are a) highly uncertain and b) do not scale appropriately for the three different models.

The first column of Figure 3 shows the data worth results of the complex benchmark model. For the storage prediction $S_{917}$ (Figure 3a), the results suggest that the groundwater well called Conders No.2, labeled “2,” has the highest data worth. This is consistent with the flow paths and head gradients in the model, which suggest that the river recharge in this part of the model domain is the major driving force behind changes in aquifer storage. In regard to the $H_{WR}$ prediction at the Wratts Rd. well (Figure 3b), the complex model estimates that the measurement data series with the highest data worth is, to no surprise, the Wratts Rd. groundwater heads.

The second and third column present the data worth estimates of the SM$_{2,POD}$ and SM$_{1,sMm}$ surrogates, respectively. The SM$_{2,POD}$ surrogate identifies the same groundwater wells of highest data worth as the CM benchmark for both model predictions clearly but marginally. This suggests it is generally capable of evaluating the existing monitoring network for these predictions, despite the high simplification error of the SM$_{2,POD}$ surrogate for the $H_{WR}$ prediction (Figure 3b) shown in Figure 2. Some realizations identify other wells as most important, mainly well number 5, Murphys Rd. This shows a high dependency of the SM$_{2,POD}$ surrogate on the inclusion of the 100 parameterizations in the results to correctly estimate existing data worth, as a single or very few realizations could lead to erroneous results otherwise. Tests with the other four model predictions (not shown here) confirm these results. The SM$_{1,sMm}$ surrogate reproduces the data worth results of the CM model for both predictions with high agreement between the different realizations, showing its robustness in regards to parameter uncertainty. Overall, the results suggest that both surrogate models are capable of correctly reproducing the data worth of the existing monitoring network, regardless of model prediction of interest and simplification error associated with the surrogate’s general predictive performance.

**Worth of “Additional” Data**

Figure 4 shows the mean worth of “additional” groundwater head measurements evaluated for locations on a grid (represented as dots in the graph) for six different model predictions. For this analysis, “additional” groundwater head data at the grid points is defined as unknown measurements in a single non-existing well assumed to be measured continuously during the calibration period.

The first column of Figure 4 shows the data worth results of the complex benchmark model. For the storage prediction $S_{917}$ (Figure 4a), the results suggest that the groundwater well called Conders No.2, labeled “2,” has the highest data worth. This is consistent with the flow paths and head gradients in the model, which suggest that the river recharge in this part of the model domain is the major driving force behind changes in aquifer storage. In regard to the $H_{WR}$ prediction at the Wratts Rd. well (Figure 4b), the complex model estimates that the measurement data series with the highest data worth is, to no surprise, the Wratts Rd. groundwater heads.
Figure 4. Mean worth of “additional” data (black dots) for CM, SM2,POD and SM1,sMm models regarding predictions of: (a) cumulative aquifer storage at \( t = 917 \) days \( (S_{917})\), (b) minimum groundwater head at Wratts Rd. well \( (H_{WR})\), (c) minimum groundwater head at potential new well 15 \( (H_{N15})\), (d) mean total river-groundwater exchange flux \( (Q_{ex})\), (e) mean river-groundwater exchange flux between gauging stations Wratts Rd. and SH1 \( (Q_{ex3})\), and (f) mean Southern Streams spring flux \( (Q_{sp3})\).

analysis of existing data in the previous chapter. When predicting the head \( H_{N15} \) (Figure 4c), the accompanying well itself is identified to have the highest data worth. For the river-groundwater exchange fluxes in the whole model area, \( Q_{ex} \) (Figure 4d), data worth of “additional” groundwater measurements is high in the west of the model domain. This is probably important for the overall groundwater head gradient in the model. Predicting the exchange flux between the Wratts Rd. and SH1 gauging stations, \( Q_{ex3} \) (Figure 4e), data worth is high in the area under the river between these two gauging stations. Finally, the highest data worth regarding the prediction of the flux into the Southern Streams, \( Q_{sp3} \) (Figure 4f), is in the south-eastern part of the model domain, where the main infiltration into these streams takes place.

Overall, the areas of high data worth of the CM are also identified by the SM2,POD surrogate. While the areas of high data worth for the first three predictions (Figure 4a-4c) are larger for the SM2,POD surrogate than the CM model, they are narrower for the other three predictions (Figure 4d-4f). For four of the six chosen predictions (Figure 4b and 4d-4f), estimates of the simplification error are shown in Figure 2. As discussed earlier, three of them (Figure 4d-4f) exhibit high simplification errors and large biases, indicating difficulties to properly reproduce the predictions by the SM2,POD surrogate. This might be a reason for the narrower zones of high data worth as estimated for these predictions by the SM2,POD surrogate. Nonetheless, high simplification errors and strong bias does not seem to effect the general ability of the SM2,POD surrogate to identify zones of high data worth in regards to “additional” groundwater head measurements for all predictions.

The third column of Figure 4 presents the data worth results of the SM1,sMm surrogate. There is some overlap between the areas of high data worth identified by the SM1,sMm surrogate and the CM benchmark for the storage prediction \( S_{917} \) (Figure 4a) and groundwater head predictions \( H_{WR} \) and \( H_{N15} \) (Figure 4b and 4c), as well as for the spring flow prediction \( Q_{sp3} \) (Figure 4f). The SM1,sMm model fails to accurately reproduce the areas of high data worth found by the CM model for...
Figure 5. 90% quantiles of worth of “additional” data (black dots) for CM, SM2,POD and SM1,sMm models regarding predictions of: (a) mean Southern Streams spring flux (Qsp3), and (b) minimum groundwater head at Wratts Rd. well (HWR).

River-groundwater exchange flux predictions Qex and Qex3 (Figure 4d and 4e), though, and identifies a larger area of high data worth for prediction Qsp3 (Figure 4f) that can not be found in the data worth map of the benchmark model. This is despite the small simplification errors of the SM1,sMm model for the three model predictions HWR, Qex and Qex3 (Figure 2). This suggests that worth of “additional” data generated with the SM1,sMm surrogate is not generally trustworthy. As the idea behind using a surrogate for data worth analysis is to reduce the computational burden of the complex model while emulating its data worth, we found the SM1,sMm model unable for this task in regard to “additional” data.

Figure 5 shows the 90% quantile maps of data worth (calculated from the 100 different parameterizations) for the two model predictions HWR and Qsp3. This illuminates the necessity of the robust data worth analysis method which takes parametric uncertainty into account. For prediction Qsp3 (Figure 5a), the area of high data worth variance coincides with the area of high data worth and is similar between the CM model and the surrogates. This suggests that for this prediction, fewer model realizations (or just the single one of the original calibrated parameter set) would have been enough to identify the area of high data worth for both the benchmark model and the surrogates. This suggests that for this prediction, fewer model realizations (or just the single one of the original calibrated parameter set) would have been enough to identify the area of high data worth for both the benchmark model and the surrogates. However, using a surrogate for data worth analysis is to reduce the computational burden of the complex model while emulating its data worth, we found the SM1,sMm model unable for this task in regard to “additional” data.

Figure 5b shows the different case for prediction HWR: while the CM model again identifies the area of high data worth in the mapping of 90% data worth quintile (suggesting that much fewer runs would suffice), the surrogate models do not. Despite identifying areas of high data worth variances at very different places, both SM2,POD and SM1,sMm models were able to correctly identify the area of high data worth for this prediction HWR (Figure 4b). This clearly demonstrates the necessity of the robust data worth analysis method to account for parametric uncertainty when using surrogate models for such predictions. We identified this for several predictions, especially for the SM2,POD model (results not shown), suggesting that while we deem the SM2,POD model a suitable surrogate for data worth analysis for “additional” data, this only holds true if taking parametric uncertainty into account. This was seen in similar fashion in the results of the existing data worth analysis above. As this can only be known by comparative analysis with the benchmark, we suggest to perform the robust data worth analysis with surrogate models for all predictions, especially where POD is applied.

Worth of “Parametric” Data

Finally, we analyzed the ability of the surrogate models to reproduce the data worth findings of the CM model in regard to parameter knowledge, that is “parametric” data. Figure 6 shows the results for three different predictions, only mapping the most influential parameters of each model.

For the prediction of aquifer storage S917 in Figure 6a, the CM model identifies an area of specific yield in layer 1 as influential parameters, which is plausible because specific yield is the key parameter in determining aquifer storage in the unconfined groundwater system. The CM benchmark parameters with the highest data worth for the flux prediction Qex3 (Figure 6b) are the river-bed conductivities in the corresponding model area. Similarly, the drain-bed conductivity of the corresponding spring boundary condition has the highest data worth regarding the mean flux through the Southern Streams Qsp3 (Figure 6c) for the benchmark model.

The second column of Figure 6 shows the results of “parametric” data worth for the SM2,POD surrogate. It is able to identify the correct parameter group (specific yield in layer 1) and the general area as the benchmark model of high data worth regarding the prediction of aquifer storage S917 (Figure 6a). The same can be seen for the predictions Qex3 (Figure 6b) and Qsp3 (Figure 6c), where the same model parameters have high data worth for the SM2,POD surrogate and the benchmark. This suggests that the SM2,POD surrogate to estimate “parametric” data worth likely stems from the fact that it keeps the full parameter set of the CM benchmark model:
Figure 6. Worth of “parametric” data for CM, SM$_{2,POD}$ and SM$_{1,sMm}$ models, either shown as maps of mean data worth (where spatially distributed parameters are most sensitive) or box plots of mean and variance (only most sensitive parameters shown). Rows delineate predictions of: (a) cumulative aquifer storage at $t = 917$ days ($S_{917}$), (b) mean river-groundwater exchange flux between gauging stations Wratts Rd. and SH1 ($Q_{ex3}$), and (c) mean Southern Streams spring flux ($Q_{sp3}$).

while the equation system of the numerical model is solved in a subspace, the underlying sensitivities between model parameters and outputs are kept. Again, the SM$_{2,POD}$ model data worth results were accompanied by a higher variance, though, which requires the robust data worth analysis method to handle parametric uncertainty. Furthermore, the application is again accurate despite the high simplification errors associated with the predictions for the SM$_{2,POD}$ surrogate, as discussed earlier.

The SM$_{1,sMm}$ surrogate, on the other hand, uses aggregates of the spatially distributed parameter fields of hydraulic conductivity, specific yield and river bed conductivity as presented earlier in Table 1. Where the benchmark estimated a zone of specific yield in layer 1 with high data worth for the prediction $S_{917}$ (Figure 6a), there are no spatially distributed parameter fields in the SM$_{1,sMm}$ surrogate. Nonetheless, it still has a parameter that encapsulates the same process: $S_{SM,y,1}^M$, the specific yield of layer 1 in the SM$_{1,sMm}$ model. Figure 6a shows that the data worth analysis identifies $S_{SM,y,2}^M$, the specific yield of layer 2, as the (only) influential parameter for this prediction, though. This could be through changes in the vertical location of the water table in the surrogate compared to the CM model, but the vertical geometry is unchanged and the overall errors of the surrogate’s water table are not high enough to shift it into the second model layer. The river bed conductivities of the benchmark showing high data worth for the prediction of the exchange flux $Q_{ex3}$ are aggregated into a single parameter between the two gauging stations in the SM$_{1,sMm}$ surrogate, $K_{SM,R,4}^M$. While it shows some data worth in the estimation, other parameters, like the hydraulic conductivity of zone 4 and a drain-bed conductivity, exhibit high data worth instead. We could not come to a hydrogeological conclusion to why these parameters should inform the prediction of river-groundwater exchange $Q_{ex3}$. The same can be said for the prediction $Q_{sp3}$, where the SM$_{1,sMm}$ surrogate again identifies parameters that do not correspond to the CM benchmark parameters of high data worth and are not logically explainable either.

Generally, it seems plausible that the main hydrogeological parameters, hydraulic conductivity and specific yield, exhibit high worth in the SM$_{1,sMm}$ surrogate due to how few and large they are in the model. Nonetheless, the surrogate fails to identify the worth of other parameters that directly influence the prediction in question, such as drain-bed conductivities when predicting the flux through the drain. A possible explanation for the inability of the SM$_{1,sMm}$ surrogate to estimate “parametric” data worth despite being equipped with consolidated model parameters representing all features of the CM benchmark could be found in the analysis of the so-called parameter surrogacy. Previous work by Gosses and Wöhling (2019) and Gosses (2020) included an analysis of parameter surrogacy for the SM$_{1,sMm}$ model. Highly abbreviated, parameter surrogacy can be defined as the characteristic of surrogate models to incorporate the correlations between model outputs and complex model parameters into different simple model parameters to compensate for their absence. Our findings in Gosses and Wöhling (2019) showed high levels of parameter surrogacy for the SM$_{1,sMm}$ model. In short, all SM$_{1,sMm}$ parameters are contributed to by some combination of
hydraulic conductivity and river-bed parameters, regardless of type and location. Furthermore, certain localized SM1,SMm parameters, such as the river-bed conductivity between gauging stations Rock Ferry and SH6, \( K_{SM1} \), do not show significant contributions from the corresponding CM parameter counterparts. This entanglement of CM parameters in the SM1,SMm surrogate parameters through parameter surrogacy explains its inability to reproduce worth of “parametric” data.

**Discussion**

In this study, we tested two very different surrogate models in their ability to be applied for analysis of worth of existing, “additional” and “parametric” data using a fast first-order second-moment method while taking into account parametric uncertainty through null-space Monte Carlo methods.

We chose a complex, physically based groundwater model as a benchmark for the study. This model was still simple enough to perform the robust data worth analysis presented here to ensure it functions as a comparison baseline for both surrogates.

The structural simplification surrogate could accurately reproduce the worth of existing data, but not of “additional” or “parametric” data. Its inability to estimate worth of “additional” data is likely caused by its simplification regarding certain processes and model structure, so it might be acceptable for predictions independent of these simplifications. “Parametric” data worth cannot be estimated by this surrogate due to its high simplification in parameter space. This could be different for a structural simplification surrogate keeping the original parameter space, though the savings in model run time would be smaller for such a surrogate. This would counteract the major advantage of this surrogate, as the run time reduction associated with it is mainly gained through the fast calculation of the Jacobian matrix due to the small number of parameters. As such, structural simplification is the easiest method to implement, but has to be applied carefully regarding predictions of interest and their dependence on system states and parameters, or it is severely hampered in its ability to estimate data worth.

The POD surrogate accurately reproduces the data worth of the complex model for all three types of data, irrespective of model prediction and associated simplification error. This shows that while the predictive variance might be larger in certain cases, it could still be applied to reliably estimate worth of data for such model predictions. This arguably stems from the fact that in retains the full complex model parameterization and spatially distributed process representation. While certain predictions are highly biased and show high simplification errors due to shortcomings of the POD model’s linearization, the overall correlations between those predictions and the system states in the model stay intact. The high variance of most of its data worth estimates suggests that parametric uncertainty has to be accounted for by the robust data worth analysis presented here for successful application. While the task of repeated computation of the Jacobian matrix necessary for this is somewhat time consuming because the original, big parameter space is kept for this surrogate, this is counteracted by the short run time of single model runs achieved through the POD method. Thus, the POD method’s good performance for data worth estimation is only hampered by the difficult setup of POD itself.

In general, our results demonstrate clearly that surrogate models have a high potential to be applied for computationally efficient, yet robust data worth analysis. Correct estimates of data worth depend on the chosen surrogate, the prediction of interest and the type of data that is analyzed. Analysis of the worth of the existing data seems to be feasible with quite constrained surrogates that may be less accurate for other data types. This suggests that potential shortcomings of the surrogates are masked by their heightened sensitivity toward the calibration data set due to the parameter estimation process. Worth of “additional,” yet unknown head measurements calculated by surrogates seems to be dependent on keeping the spatial dimensions of the complex model, at least for some predictions. “Parametric” data worth estimation necessitates a full preservation of the complex model’s parameter space in the surrogate. These requirements seem to be more important than the overall trustworthiness of the surrogate prediction, as the simplification error analysis suggests. Results generated via first-order second-moment methods and surrogates have to be taken with caution, though, as the high variance of many data worth estimates by the surrogates suggests that parametric uncertainty has to be taken into account. For this, the proposed method to combine first-order second-moment uncertainty estimation methods with multiple parameter fields efficiently generated by, for example, null-space Monte Carlo, seems to be sufficient. Furthermore, absolute values of predictive uncertainty and data worth produced by the first-order second-moment methods should not be evaluated in themselves, but the relative changes and thus sensitivities of model predictions toward data are reasonable estimates.

**Summary and Conclusions**

In this work, we applied a first-order second-moment data worth analysis with two different surrogate models and a complex benchmark for comparison. To address the influence of parameter non-uniqueness, we utilized multiple calibrated parameterizations in the analysis which were generated efficiently with null-space Monte Carlo methods. The results of our data worth analysis with surrogate models are summarized as follows (Table 3):

- The simplified MODFLOW model reproduced worth of existing data correctly, but is limited in its use for “additional” data worth analysis due to its structural simplification. The parametric simplification renders it completely useless for the assessment of “parametric” data worth.
The POD surrogate accurately reproduced worth of existing, “additional” and “parametric” data of the complex model for all prediction types considered in this study. The high variance associated with many of its results suggest that this achievement requires the application of the robust data worth method presented here which takes parametric uncertainty into account.

- Both surrogate models achieve significant run time reduction, either due to the short run time of single model runs (POD surrogate) or by the effect of parameter reduction on the estimation of model sensitivities (simplified MODFLOW surrogate).
- The ability of the surrogate models to reproduce data worth is independent of the simplification error associated with the prediction of interest.

From these results, we can conclude that surrogate models are well suited for the task of estimating data worth analysis. The dependence of “additional” and “parametric” data worth on maintaining the closest possible parameterization, spatial dimensions and process representation of the complex model suggests that projection-based surrogates are more suited for these tasks compared to spatially and parametrically simplified models. These results, together with the higher bias and simplification error of many of the POD surrogate’s predictions in comparison to the simplified MODFLOW surrogate.

The POD surrogate accurately reproduced worth of the categorization of suitable surrogates for data worth estimation, that is, which type of surrogate is appropriate for which type of prediction and data, has to be gained to generate guidelines, or even automated criteria, for the choice of best possible surrogate for the contrived task.

### Acknowledgments
This work was funded by the German Research Foundation (grant WO 1781/1-1). The author would like to thank the Marlborough District Council for providing the data for this study, as well as Michael Fienen and one anonymous reviewer for the insightful comments on the initial version of this paper. Open Access funding enabled and organized by Projekt DEAL.

### Authors’ Note
The authors do not have any conflicts of interest or financial disclosures to report.

### Appendix
This appendix holds a short overview of the model simplification method “Proper Orthogonal Decomposition” (POD). More thorough presentations of the method’s application in groundwater can be found in Vermeulen et al. (2004a, 2004b), McPhee and Yeh (2008), Pasetto et al. (2013), Boyce and Yeh (2014), Stanko et al. (2016) and Gosses et al. (2018), among others.

POD simplifies a linear, physically distributed groundwater model by subspace projection. A linear groundwater model with implicit time-stepping can be formulated as:

$$A\mathbf{d}^{t+1} = \mathbf{b},$$

where $\mathbf{A}$ holds a combination of the stiffness matrix, holding the hydraulic conductivity, and the mass matrix, containing the storage. $\mathbf{d}$ is the vector of groundwater heads in the model cells, usually named $\mathbf{h}$ in literature, but renamed here to avoid overlap, and the vector $\mathbf{b}$ holds boundary conditions and external forcings. Note that the model parameters, $\mathbf{k}$, are usually part of the calculation of the matrix $\mathbf{A}$. Model run time is majorly affected by the number of model cells, $n_m$, as the main computational task of solving the equation system is inversion of the matrix $\mathbf{A}$, which has the size $\mathbf{A} \in \mathbb{R}^{n_m \times n_m}$.

Model reduction by POD is obtained by subspace projection of Equation (7). The generation of this subspace is undertaken by singular value decomposition of a so-called snapshot matrix. Here, a snapshot is a solution of first-order second-moment methods can be extended with efficiently generated multiple parameter fields to safeguard their estimations against bias caused by parameter non-uniqueness.

Many open questions remain in the research of surrogate models are well suited for the task of estimating data worth analysis. The dependence of “additional” and “parametric” data worth on maintaining the closest possible parameterization, spatial dimensions and process representation of the complex model suggests that projection-based surrogates are more suited for these tasks compared to spatially and parametrically simplified models. These results, together with the higher bias and simplification error of many of the POD surrogate’s predictions in comparison to the simplified MODFLOW surrogate.

| Data worth reproduction | Structural and parametrical simplification | Proper orthogonal decomposition (POD) |
|-------------------------|-------------------------------------------|-------------------------------------|
| “Additional”            | Some                                      | All                                 |
| “Parametric”            | No                                        | All                                 |
| Influence of simplification error | Not apparent   | Not apparent                        |

### Table 3
Summary of Surrogate Data Worth Performance

| Data worth reproduction | Structural and parametrical simplification | Proper orthogonal decomposition (POD) |
|-------------------------|-------------------------------------------|-------------------------------------|
| “Additional”            | Some                                      | All                                 |
| “Parametric”            | No                                        | All                                 |
| Influence of simplification error | Not apparent   | Not apparent                        |

The authors do not have any conflicts of interest or financial disclosures to report.
matrix to singular value decomposition, $H_s = USV^T$, gives the matrix of left singular vectors $U \in \mathbb{R}^{m \times m}$, among other matrices $S$ and $V$. The singular vectors in $U$ now hold the variability of $H_s$ in descending order. Now, a projection matrix can be defined as:

$$P = [u_1, u_2, \ldots, u_{n_r}], \quad (9)$$

which holds the $n_r$ first columns of $U$. The size of the surrogate system, $n_r$, can be defined from the singular values in $S$, which rank the amount of variability held in each singular vector $u$. This amount, called relative retained variance, is usually chosen to be $>99\%$. To generate the final equation system of the POD surrogate, the projection matrix is multiplied with the linear equation system of the groundwater model, Equation 7, to get:

$$P^T A P r^{t+1} = P^T b, \quad (10)$$

from which the original solution of groundwater heads is calculated for each timestep as $d \equiv P r$. As $n_r \ll n_m$, Equation 10 can be solved much faster than Equation 7.

It is of note that the POD surrogate still utilizes the full parameterization of the original model due to its dependence of $A$, which holds the model parameters. Thus, tasks as calculation of the Jacobian matrix or full model calibration can be undertaken with the surrogate on those model parameters. As $P$ is dependent on full model solutions through $H_s$, deviation of the model parameters from the parameter set chosen to generate $H_s$ potentially degrades the quality of the surrogate.

References

Adamowski, J., and H.F. Chan. 2011. A wavelet neural network conjunction model for groundwater level forecasting. *Journal of Hydrology* 407, no. 1: 28–40. https://doi.org/10.1016/j.jhydrol.2011.06.013

Asher, M.J., B.F.W. Croke, A.J. Jakeman, and L.J.M. Peeters. 2015. A review of surrogate models and their application to groundwater modeling. *Water Resources Research* 51, no. 8: 5957–5973. https://doi.org/10.1002/2015WR1016967

Beven, K., and A. Binley. 1992. The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes* 6, no. 3: 279–298. https://doi.org/10.1002/hyp.36000305

Boyce, S.E., and W.W.-G. Yeh. 2014. Parameter-independent model reduction of transient groundwater flow models: Application to inverse problems. *Advances in Water Resources* 69: 168–180. https://doi.org/10.1016/j.adwres.2014.04.009

Burrows, W., and J. Doherty. 2016. Gradient-based model calibration with proxy-model assistance. *Journal of Hydrology* 533: 114–127. https://doi.org/10.1016/j.jhydrol.2015.11.033

Cirpka, O.A., C.M. Bürger, W. Nowak, and M. Finkel. 2004. Uncertainty and data worth analysis for the hydraulic design of funnel-and-gate systems in heterogeneous aquifers. *Water Resources Research* 40, no. 11: W11502. https://doi.org/10.1029/2004WR003352

Daliakopoulos, I., P. Coulibaly, and I.K. Tsanis. 2005. Groundwater level forecasting using artificial neural networks. *Journal of Hydrology* 309: 229–240. https://doi.org/10.1016/j.jhydrol.2004.12.001

Dausman, A.M., J. Doherty, C.D. Langevin, and M.C. Sukop. 2010. Quantifying data worth toward reducing predictive uncertainty. *Groundwater* 48, no. 5: 729–740. https://doi.org/10.1111/j.1745-6584.2010.00679.x

Doherty, J. 2016. PEST: Model-Independent Parameter Estimation – User Manual, 6th ed. Brisbane, Australia: Watermark Numerical Computing.

Doherty, J., and S. Christensen. 2011. Use of paired simple and complex models to reduce predictive bias and quantify uncertainty. *Water Resources Research* 47, no. 12: W12534. https://doi.org/10.1029/2011WR010763

Erickson, M., A. Mayer, and J. Horn. 2002. Multi-objective optimal design of groundwater remediation systems: Application of the niched pareto genetic algorithm (NPGA). *Advances in Water Resources* 25, no. 1: 51–65. https://doi.org/10.1016/S0309-1708(01)00020-3

Fan, Y., W. Lu, T. Miao, J. Li, and J. Lin. 2020. Multiobjective optimization of the groundwater exploitation layout in coastal areas based on multiple surrogate models. *Environmental Science and Pollution Research* 27: 19561–19576. https://doi.org/10.1007/s11356-020-08367-2

Feyen, L., and S.M. Gorelick. 2005. Framework to evaluate the worth of hydraulic conductivity data for optimal groundwater resources management in ecologically sensitive areas. *Water Resources Research* 41, no. 3: W03019. https://doi.org/10.1029/2003WR002901

Fienen, M., J. Doherty, R. Hunt and H. Reeves. 2010. Using prediction uncertainty analysis to design hydrologic monitoring networks: Example applications from the Great Lakes water availability pilot project. U.S. Geological Survey Scientific Investigations Report 5159. https://pubs.usgs.gov/sir/2010/5159 (accessed October 2, 2020)

Fu, J., and I.J. Gómez-Hernández. 2009. Uncertainty assessment and data worth in groundwater flow and mass transport modeling using a blocking Markov chain Monte Carlo method. *Journal of Hydrology* 364, no. 3: 328–341. https://doi.org/10.1016/j.jhydrol.2008.11.014

Gosses, M. 2020. Model-based data worth analysis for groundwater systems with the use of surrogate models. Ph.D. thesis, Technische Universität Dresden, Dresden, Germany. https://nbn-resolving.org/urn:nbn:de:bsz:14-qucosa2-729241 (accessed December 1, 2020)

Gosses, M., and T. Wöhling. 2019. Simplification error analysis for groundwater predictions with reduced order models. *Advances in Water Resources* 125: 41–56. https://doi.org/10.1016/j.adwres.2019.01.006

Gosses, M., W. Nowak, and T. Wöhling. 2018. Explicit treatment for Dirichlet, Neumann and Cauchy boundary conditions in POD-based reduction of groundwater models. *Advances in Water Resources* 115: 160–171. https://doi.org/10.1016/j.adwres.2018.03.011

von Gunten, D., T. Wöhling, C. Haslauer, D. Merchan, J. Causapé, and O. Cirpka. 2014. Efficient calibration of a distributed pde-based hydrological model using grid coarsening. *Journal of Hydrology* 519: 3290–3304. https://doi.org/10.1016/j.jhydrol.2014.10.025

Khu, S.-T., and M.G.F. Werner. 2003. Reduction of Monte-Carlo simulations runs for uncertainty estimation in hydrological modelling. *Hydrology and Earth System Sciences Discussions* 7, no. 5: 680–692.

Knowling, M.J., J.T. White, C.R. Moore, P. Rakowski, and K. Hayley. 2020. On the assimilation of environmental tracer observations for model-based decision support. *Hydrology and Earth System Sciences* 24, no. 4: 1677–1689. https://doi.org/10.5194/hess-24-1677-2020

Matott, L.S., J.E. Babendreier, and S.T. Purucker. 2009. Evaluating uncertainty in integrated environmental models: A review of concepts and tools. *Water Resources Research* 45, no. 6: W06421. https://doi.org/10.1029/2008WR007301

McPhee, J., and W.W.-G. Yeh. 2008. Groundwater management using model reduction via empirical orthogonal functions.
Sreekanth, J., D. Gladish, D. Gonzalez, D. Pagendam, T. Sbai, M.A. 2019. Well rate and placement for optimal groundwater remediation of a coastal aquifer system of the Venice lagoon. Engineering Applications of Artificial Intelligence 25, no. 8: 1670–1676. https://doi.org/10.1016/j.engappai.2012.02.009

Ushijima, T.T., and W.W.-G. Yeh. 2015. Experimental design for estimating unknown hydraulic conductivity in an aquifer using a genetic algorithm and reduced order model. Advances in Water Resources 86: 193–208. https://doi.org/10.1016/j.advwatres.2015.09.029

Vermeulen, P., A. Heemink, and C. te Stroet. 2004a. Reduced models for linear groundwater flow models using empirical orthogonal functions. Advances in Water Resources 27, no. 1: 57–69. https://doi.org/10.1016/j.advwatres.2003.09.008

Vermeulen, P., A. Heemink, and C. te Stroet. 2004b. Low-dimensional modelling of numerical groundwater flow. Hydrological Processes 18, no. 8: 1487–1504. https://doi.org/10.1002/hyp.1424

Watson, T.A., J.E. Doherty, and S. Christensen. 2013. Parameter and predictive outcomes of model simplification. Water Resources Research 49, no. 7: 3952–3977. https://doi.org/10.1002/wrcr.20145

Winston, R. 2009. ModelMuse – A graphical user interface for MODFLOW-2005 and PHAST. U.S. Geological Survey Techniques and Methods 6-A29. https://pubs.usgs.gov/tm/tm6A29/ (accessed June 25, 2020)

Winton, C., J. Pettway, C. Kelley, S. Howington, and O.J. Eslinger. 2011. Application of proper orthogonal decomposition (POD) to inverse problems in saturated groundwater flow. Advances in Water Resources 34, no. 12: 1519–1526. https://doi.org/10.1016/j.advwatres.2011.09.007

Wöhlung, T., S.R. Wilson, V. Wadsworth, and P. Davidson. 2020. Detecting the cause of change using uncertain data: Natural and anthropogenic factors contributing to declining groundwater levels and flows of the Wairau Plain Aquifer, New Zealand. Journal of Hydrology: Regional Studies 31: 100715. https://doi.org/10.1016/j.jhyps.2020.100715

Wöhlung, T., M.J. Gosses, S.R. Wilson, and P. Davidson. 2018. Quantifying river-groundwater interactions of New Zealand’s gravel-bed rivers: The Wairau plain. Groundwater 56, no. 4: 647–666. https://doi.org/10.1111/gwat.12625

Wöhlung, T., A. Geiges, and W. Nowak. 2016. Optimal design of multitype groundwater monitoring networks using easily accessible tools. Groundwater 54, no. 6: 861–870. https://doi.org/10.1111/gwat.12430

Yin, J., and F.T.-C. Tsai. 2020. Bayesian set pair analysis and machine learning based ensemble surrogates for optimal multi-aquifer system remediation design. Journal of Hydrology 580: 124280. https://doi.org/10.1016/j.jhydrol.2019.124280

Yoon, H., S.-C. Jun, Y. Hyun, G.-O. Bae, and K.-K. Lee. 2011. A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer. Journal of Hydrology 396, no. 1: 128–138. https://doi.org/10.1016/j.jhydrol.2010.11.002