# Impact of model complexity and multi-scale data integration on the estimation of hydrogeological parameters in a dual-porosity aquifer

Elena Tamayo-Mas[1], Marco Bianchi[1], and Majdi Mansour[1]

[1]British Geological Survey, Environmental Science Centre, Nicker Hill, Keyworth, Nottingham, NG12 5GG, United Kingdom

Corresponding author: Elena Tamayo-Mas, British Geological Survey, Environmental Science Centre, Nicker Hill, Keyworth, Nottingham, NG12 5GG, United Kingdom, email: elena@bgs.ac.uk.

**Abstract.** This study investigates the impact of model complexity and multi-scale prior hydrogeological data on the interpretation of pumping test data in a dual-porosity aquifer (the Chalk aquifer in UK). In order to characterize the hydrogeological properties, different approaches ranging from a traditional analytical solution (Theis approach) to more sophisticated numerical models with automatically calibrated input parameters are applied. Comparisons of results from the different approaches show that neither traditional analytical solutions nor a numerical model assuming a homogenous and isotropic aquifer can adequately explain the observed drawdowns. A better reproduction of the observed drawdowns in all seven monitoring locations is instead achieved when medium and local scale prior information about the vertical hydraulic conductivity distribution (*K*) is used to constrain the model calibration process. In particular, the integration of medium scale vertical *K* variations based on flowmeter measurements lead to an improvement in the goodness-of-fit of the simulated drawdowns of about 30%. Further improvements (up to 70%) were observed when a simple upscaling approach was used to integrate small-scale *K* data to constrain the automatic calibration process of the numerical model. Although our analysis focuses on a specific case study, these results provide insights about the representativeness of the estimates of hydrogeological properties based on different interpretations of pumping test data, and promote the integration of multi-scale data for the characterization of heterogeneous aquifers in complex hydrogeological settings.

## 1 Introduction

Quantitative characterization of physical and chemical properties of aquifers is a critical task for groundwater investigations (e.g. Fogg et al., 1998; Gelhar and Axness, 1983; Koltermann and Gorelick, 1996; Poeter and Gaylord, 1990). The reliability of modelling tools supporting decisions about the development, management, and protection of groundwater resources depends on accurate estimations of properties such as hydraulic conductivity ($K$), transmissivity ($T=Kb$, where $b$ is the aquifer thickness), and storativity ($S$), which largely control groundwater flow and solute transport in geological media.

Obtaining representative values for these properties is not a trivial task. One challenge is the intrinsic heterogeneity of geological media and the consequential variability of the hydrogeological properties, which can be of several orders of magnitude within the same aquifer system (e.g. Bohling et al., 2016; Fogg et al., 1998; Oehlmann et al., 2013; Williams et al., 2006). Moreover, because of the multiscale nature of geological media (Neuman and Di Federico, 2003), estimated effective values strongly depend on the volume of the aquifer investigated – also known as the support scale or support volume – and hence on the measurement method. For instance, it has been observed in a variety of aquifers that measured $K$ values tend to increase with the support volume (Martinez-Landa and Carrera, 2005; Odén and Niemi, 2006; Rovey and Cherkauer, 1995; Schulze-Makuch et al., 1999). A further challenge is that estimated values are usually obtained from the solution of an inverse problem in which the objective is to minimize the error between measured values of the state variables (i.e. hydraulic heads, pressures, or drawdowns) and the corresponding predictions from a model describing fluid flow in geological media. Once an optimal set of input values is found to satisfy a certain goodness-of-fit criterion, the model is considered to be calibrated and these values are then considered representative of the hydrogeological properties in the system of interest.

However, as pointed out by several studies in the past decades (reviews by Carrera et al., 2005; Oliver and Chen, 2011; Zhou et al., 2014), in most hydrogeological applications this solution is not unique and the inverse problem is ill-posed. This issue arises especially when the number of observation data is small compared to the number of parameters to evaluate, or when the outputs from the model are not sensitive to certain input parameters.

Effective values for the hydrogeological properties are generally determined from the analysis of the results of pumping tests (Sanchez-Vila et al., 2006). The support volume of these tests can vary significantly according to their duration, the length of the screened section of the pumping borehole compared to the aquifer thickness, the adopted pumping rate, as well as the aquifer permeability. Traditionally, the approach for the interpretation of observed time-drawdown data and the estimation of the hydrogeological properties consist in the calibration of analytical solutions of the 1-D partial differential equation describing transient radial flow in a homogeneous porous media (e.g. Bear, 2007; Domenico and Schwartz, 1997; Fetter, 2000). Commonly used solutions (e.g. Theis, 1935; Cooper and Jacob, 1946) assume a fully penetrating borehole (i.e. screen length equal to the aquifer thickness) in a homogeneous, isotropic and confined aquifer. Several other analytical solutions have been introduced over the years to cover a wide variety of hydrogeological and boundary conditions. Reviews of the different solutions are provided by Kruseman and de Ridder (1990) and more recently by Yeh and Chang (2013).

Notwithstanding the wide range of available analytical solutions and their widespread use for hydrogeological characterization, their application may become questionable for the interpretation of pumping tests in hydrogeological settings characterized by high heterogeneity and complex boundary conditions such as in the presence of groundwater – surface water interactions, aquifer recharge, complex aquifer geometry, variable pumping rates, or multiple

boreholes interference. In these situations, the simplifying assumptions made to derive these solutions are usually not adequately representative of the system of interest. A careless application of analytical solutions for aquifer characterization in these settings may introduce significant systematic errors in the estimated values of the hydrogeological properties, as shown for other characterization methods (e.g. Beckie and Harvey, 2002; Bianchi, 2017). One main limiting factor is the assumption of homogeneity, which in complex and heterogeneous hydrogeological settings is often contradicted by the drawdown data observed at different locations. These data, once interpreted according to an analytical solution, provide in fact a range of values for the hydrogeological properties that is inconsistent with the homogeneity assumption. These inconsistencies tend to be more apparent from the interpretation of the early time data while later time data provide more homogenous results because of their larger support scale (Meier et al., 1998; Sánchez-Vila et al., 1999).

Addressing limitations inherent analytical solutions to improve the match between observed and simulated time-drawdown curves requires the implementation of numerical models (e.g. Mansour et al., 2011; Raghavan, 2004; Schad and Teutsch, 1994; Thorbjarnarson et al., 1998). Pumping tests have been simulated with finite difference (Barrash and Dougherty, 1997; Cheng and Chen, 2007; Halford and Yobbi, 2006; Kaleris et al., 1995; Lee et al., 2007; Leven and Dietrich, 2006; Mohamed and Rushton, 2006; Raghavan, 2009, 2006; Schroth and Narasimhan, 1997), finite elements, and hybrid finite elements – finite difference models (Chen and Jiao, 1999; Lebbe et al., 1992). Radial flow models based on cylindrical coordinate grids are particularly appropriate for the simulation of pumping tests since they provide a more precise representation of the flow field around the pumping borehole (e.g. Singh, 2000). An example is the layered cylindrical grid numerical model developed by Mansour et al. (2011) to simulate complex time-drawdown curves impacted by concurring factors including the degree

of fracturing, simultaneous pumping from adjacent boreholes, a quarry with development depth below the water table introducing an internal flow boundary, and variable pumping rates. All these factors make traditional analytical solutions not suitable for the interpretation of pumping test data and the characterization of the hydrogeological properties. On the contrary, numerical modelling allows a more detailed conceptual understanding of the groundwater system.

While numerical models are powerful tools to characterize complex hydrogeological settings, they too require calibration of their input parameters. Calibration methods range from a simple manual trial-and-error approach to more complex and efficient automatic methods (Zhou et al., 2014). Because the number of input parameters that need to be adjusted to fit the observations is larger than for analytical solutions, the calibration of numerical models is more affected by the issue of non-uniqueness. Reducing the number of model input parameters with a simplification of a more complex numerical model and/or applying constraints to the variability of the input parameters on the basis of certain prior information are two effective strategies to improve the uniqueness of the solution (Zhou et al., 2014). However, the application of these strategies in complex hydrogeological systems presents some difficulties because of the risk of oversimplification (Raghavan et al., 2002). Regarding the use of prior information to constrain hydrogeological parameters, issues of representativeness may also arise when the supporting volume of this information is not the same as the scale of the pumping test or the scale of the numerical model used to simulate the observed time-drawdown curves. One solution is to include an upscaling approach in the calibration process (Raghavan, 2004), but only few studies have combined upscaling and inverse modelling approaches (e.g., Li et al., 2012).

The objective of this paper is to present a case study to investigate the impact of model complexity and integration of prior hydrogeological data at different scales on the

97 interpretation of a pumping test in a dual-porosity aquifer (the Chalk aquifer in UK). For this
98 purpose, different *K* and *S* values obtained from the calibration of both traditional analytical
99 solutions and radial flow numerical models with increasing complexity are compared. The
100 calibration of the numerical models is performed automatically with the parameter estimation
101 code PEST (Doherty, 2015) considering both unconstrained and constrained optimization
102 strategies to evaluate the impact of prior conditioning data. In particular, these consist of a set
103 of *K* values with a smaller vertical support volume compared to the radius of influence of the
104 pumping test. We show that a simple upscaling approach based on the ranges of these *K* data
105 is enough to improve the match between measured and simulated drawdown data. Although
106 we focus on a specific case study, our results provide valuable guidelines for the interpretation
107 of pumping test data and the characterization of hydrogeological properties in any
108 hydrogeological context.
109
110 **2 Site description and data**
111 **2.1 Hydrogeological setting**
112 The pumping test was performed in the Chalk aquifer at a location in the Pang-
113 Lambourn catchment of the Thames Basin (UK) in the southern England (Figure 1). The Chalk
114 is a major aquifer providing approximately 70% of the public water supply to the south east of
115 England (Allen et al., 1997). The Chalk is a generally productive dual-porosity aquifer due to
116 the elevated secondary porosity provided by fractures (e.g. Bloomfield et al., 1995), but
117 productivity varies with depth according to a typical non-linear trend of decrease in *K* (Owen
118 and Robinson, 1978). Factors controlling this trend include a reduction of fracture spacing and
119 aperture with depth (Bloomfield, 1996), as well as the amplification in the upper part of the
120 aquifer of fracture aperture in response to carbonate dissolution and other diagenetic processes

121 (Price, 1987). Accordingly, high *K* values in the range of 0.1 to over 100 m/day are observed
122 at shallow depths (50 – 60 m below the ground surface), especially within the range of seasonal
123 water table oscillations. Statistical analysis of data from more than two thousand pumping tests
124 indicates a median *T* value in the order of 500 $m^2$/day. For tests conducted in unconfined aquifer
125 conditions, the median value of the storage coefficient data is equal to 0.008, while it is equal
126 to about 0.001 for tests carried out in confined conditions (MacDonald and Allen, 2001).
127 The site of the pumping test was previously investigated extensively during a research
128 project aimed to improve the understanding of the hydrogeological conditions of major UK
129 aquifers (Wheater and Peach, 2004). In addition to pumping test data, geophysical logs, aquifer
130 samples, borehole images, impeller and heat-pulse flowmeter measurements, as well as packer,
131 dilution, and tracer tests data were also collected for the characterization of the aquifer
132 properties. Results from this very comprehensive investigation campaign are presented by
133 Williams et al. (2006), Butler et al. (2009), and Maurice et al. (2012). Data collected at this site
134 have also been used in a previous study showing the importance of an accurate characterization
135 of the input function in the interpretation of radially convergent tracer tests (Mathias et al.,
136 2007).
137 The interpretation of the borehole logs according to the most recent stratigraphical
138 model of the Chalk (Woods et al., 2015) indicates that the upper section of the aquifer (from 0
139 m to ≈ 55 m below ground surface) is within the Seaford Chalk formation. The deeper section
140 up to 90 m below ground surface consists of a 25 m thick horizon of Lewes Nodular Chalk
141 formation underlain by the upper New Pit Chalk formation. Groundwater flows in the
142 unconfined aquifer from north to south following the slope of the ground surface. The average
143 hydraulic gradient is about 0.001 (Williams et. al. 2006). The average water table depth is
144 around 20 m below ground surface with annual oscillations of about 7 m. Groundwater

145 recharge in the proximity of the site has been investigated by Ireson et al. (2009) through
146 numerical simulation of infiltration mechanisms, who suggest a continuous drainage of water
147 through the unsaturated zone to the water table even during drought conditions. However, given
148 the short duration of the pumping test considered in this work, unsaturated zone recharge fluxes
149 and fluid flow through the unsaturated zone were not taken into account in the numerical
150 simulations.

151 **2.2 Pumping test data**

152 For the execution of the pumping test, an open borehole, partially penetrating the
153 aquifer up to a maximum depth of 86 m below ground surface (borehole EA in Figure 1), was
154 pumped at a relatively constant rate (variations between 5520 and 6010 $m^3$/day) for about 35
155 hours. Hydraulic heads were continuously monitored (15 seconds intervals) with pressure
156 transducers in the abstraction borehole and in six monitoring boreholes. Three of these
157 boreholes (A, B and E) are 100 m deep open boreholes, whereas the other three (C, D and F)
158 are 40 m deep and each fitted with two piezometers to monitor heads at two different depths in
159 the aquifer (Table 1). Hydraulic head recovery was also monitored for about 15 hours after the
160 pump shutdown.

161 The pumping test data consist of ten time-drawdown curves covering both the
162 abstraction and the recovery phases of the test. Representative time-drawdown curves,
163 including those measured at the abstraction borehole (EA), the open borehole E, and at the two
164 piezometers in borehole F are presented in Figure 2, while the data collected in the other
165 boreholes (A, B, C, D, and E) are presented in Figures 4 – 8 and Figure 10. As shown in Figure
166 2, the measured rates of drawdown during the abstraction phase in the F1 and F2 observation
167 boreholes were similar only at early times, whereas at later times the curve for piezometer F2
168 becomes steeper (Figure 2a). The drawdown data in the abstraction well EA show a gradual

169 increase at the beginning of the pumping phase due to well storage effects. However, the time-
170 drawdown curve only slightly shows the typical S type-curve typical of unconfined condtions.
171 Different responses after cessation of pumping are also observed in the piezometers F1 and F2,
172 as well as in the open boreholes EA and E during the recovery phase (Figure 2b). In particular,
173 a slow recovery of the groundwater levels is observed in borehole E and in the shallowest of
174 the piezometers F (F2). The difference in the responses observed for piezometers F1 and F2
175 may be indicative of local vertical variations in the aquifer conductivity. Another feature that
176 can be related to vertical heterogeneities in the *K* distribution is the discrepancy between the
177 slopes of the late-time portions of the abstraction and recovery curves for each well (Rushton
178 and Chan, 1976). Because of these discrepancies, the interpretation of the abstraction and the
179 recovery data according to traditional analytical methods (e.g. Cooper and Jacob, 1946) would
180 result in two different values for the effective aquifer transmissivity for each observation
181 borehole (see also Butler et al., 2009). Another noticeable feature in the observed time-
182 drawdown curves is the increment in slope observed during the end of the abstraction phase
183 (time > 100 minutes). This behaviour can be explained by the presence of either large-scale *K*
184 variations consistent with a layered aquifer structure or a hydrogeological boundary. However,
185 the hydrogeological setting in the area of investigation does not provide elements to justify the
186 presence of such boundary.

187 **2.3 Prior hydrogeological information**

188 In this work, we use prior information about the vertical distribution of *K* to constrain
189 or regularize the automatic calibration of the numerical model for the simulation of the
190 pumping test. Among the available data, we only considered the *K* values estimated using the
191 constant head double-packer permeameter described by Price and Williams (1993), and
192 measures of horizontal flow in borehole A collected with an impeller flowmeter. Details about

193 these relatively simple and easy-to-perform hydrogeological tests can be found in Williams et
194 al. (2006), Mathias et al. (2007), and Butler et al. (2009). Here, we simply summarize the results
195 we used to complement the pumping test data.

196 Packer testing was conducted in the three open boreholes at the site (boreholes A, B,
197 and E in Figure 1). For the purpose of this study, *K* values at different elevations from these
198 boreholes were combined into one representative vertical profile of local *K* variability, covering
199 the saturated zone between 20 and 100 m below ground (Figure 3). The vertical support scale
200 of these *K* estimates, which for this type of data is given by the distance between the inflatable
201 packers used to isolate a specific aquifer interval in the borehole, is equal to about 3 meters.
202 Estimated *K* values indicate high heterogeneity with a total variation of almost 5 orders of
203 magnitude from the top to the bottom of the vertical profile. This trend can be well represented
204 by a logarithmic function, which appears to be observed in other chalk aquifers (Allen et al.,
205 1997; Nativ et al., 2003; Price and Williams, 1993). The local heterogeneity is the effect of the
206 relatively small support scale of these type of measurements, suggesting that aquifer
207 conductivity at this scale is controlled by discrete fractures. Accordingly, the Chalk formation
208 at this scale behaves hydraulically as a fractured aquifer rather than being equivalent to a porous
209 permeable rock. Despite this consideration, useful information can still be extracted from these
210 data in the form realistic bounds for the parameters considered for the calibration of numerical
211 models.

212 Impeller flowmeter measurements were collected along the open borehole A, covering
213 the depth interval between approximately 20 and 100 m below the ground surface. Inflows or
214 outflows estimations are based on recording variations in the frequency of rotation of an
215 impeller that is lowered at a constant velocity in the borehole. As shown by Butler at al. (2009),
216 the vertical profile of net upflows into the borehole is characterized by step changes

representing four layers with distinct flow regimes. This layered structure of the aquifer derived from the interpretation of flowmeter data collected in borehole A has been confirmed by the results of single borehole dilution tests conducted in boreholes A and B (Maurice et al., 2012), but these are not taken into consideration in this work. The boundaries of the four identified layers are superimposed on the packer test data in Figure 3. In particular, the analysis of the flowmeter data indicates a layer of inflows from the bottom of the borehole (100 m below ground surface) up to 83 m of depth. This deepest layer underlays a 30 m thick layer characterized by additional inflow, which is in turn overlain by a layer of outflows between 36 m and 53 m below ground. The shallowest layer, from 36 m up to the water table (about 20 m below ground) is characterized by significantly larger outflows, which is consistent with higher transmissivity values measured at shallow depths in the Chalk. Compared to the packer test data, the flowmeter measurements suggest vertical variations in $K$ in the order of tens of meters, i.e. comparable to the thickness of the different horizons. These variations are used in this work to define the thicknesses of the layers in the numerical models and therefore to constrain the spatial variability of the parameters considered for model calibration.

## 3 Methods of interpretation of the pumping test data

### 3.1 Theis analytical solution

Effective transmissivities and storage coefficients from the application of the Theis solution to the drawdowns measured during the abstraction phase in three representative boreholes are reported in Table 2. These differ from the values estimated by Butler et al. (2009), which instead are the result of the analysis of the late-time recovery data. The Theis solution assumes confined conditions in a homogeneous isotropic porous aquifer and a fully penetrating pumping abstraction borehole. Due to the hydrogeological setting, the relatively small scale

241 and duration of the pumping test, as well as the characteristics of the abstraction and monitoring
242 boreholes (i.e. partially penetrating open boreholes and piezometers), these conditions are
243 unlikely to be fully attained. More sophisticated analytical methods (e.g. Mathias and Butler,
244 2006; Moench, 2003; Neuman, 1972) could also be applied to estimate aquifer properties.
245 However, reference values from the more commonly applied Theis solution are presented here
246 for comparison with the numerical results to highlight some of the challenges regarding the
247 application of conventional approaches for the interpretation of complex pumping tests. The
248 results of this analysis are anticipated in the next paragraph and they will not be further
249 discussed in the results section.

250 As shown in Figure 4, both the early and the late time segments of the abstraction phase
251 data were considered for interpretation. Values of the hydrogeological properties derived from
252 the interpretation of the late drawdown curve are representative of a larger aquifer volume
253 compared to those derived from early times, which instead are more sensitive to local aquifer
254 heterogeneity. In fact, similar late time based transmissivity values were estimated for the three
255 analyzed boreholes, although the interpretation of data for borehole B suggests a 10% higher
256 effective transmissivity and a lower storage coefficient, see Table 2. Instead, the interpretation
257 of the early time curves resulted in a set of rather different estimates for the hydrogeological
258 properties. This variability complicates the interpretation of the pumping test analysis and in
259 particular, complicates the determination of a unique set of values that can be considered
260 representative of the actual hydraulic properties of the aquifer. For instance, if the arithmetic
261 means of the late time based values for $T$ (about 1030 $m^2$/day) and $S$ (about 0.003) are used as
262 a representative values, the Theis model leads to a result that does not match any of the
263 observed time-drawdown data. In addition, the Theis model does not provide any information
264 about the vertical hydraulic conductivity of the aquifer, a parameter that highly affects the

shape of the time-drawdown curve in unconfined aquifers at locations in proximity to the pumping well (Neuman, 1972).

### 3.2 Numerical models

#### 3.2.1 Radial flow model

Radial flow in the unconfined chalk aquifer was simulated with the object oriented code COOMPuTe (Mansour et al., 2007). This model is based on a finite difference approximation of the 3-D governing equation of flow in porous media expressed in cylindrical coordinates (Rushton, 2003):

$$\frac{K_r}{r}\frac{\partial h}{\partial r} + K_r\frac{\partial^2 h}{\partial r^2} + \frac{K_\theta}{r}\frac{\partial^2 h}{\partial \theta^2} + K_z\frac{\partial^2 h}{\partial z^2} + N = S_s\frac{\partial h}{\partial t} \quad (1)$$

where $h(r,\theta,z)$ is the hydraulic head [L] at a point at cylindrical coordinates $(r,\theta,z)$, $S_s$ is the volumetric specific storage [$L^{-1}$], $N$ is a sink-source per unit volume term that is positive for recharge and negative for withdrawal [$T^{-1}$], and $K_r, K_\theta$ and $K_z$ [$LT^{-1}$] are the components of the hydraulic conductivity tensor in the respective cylindrical coordinates directions. Because the coordinate system is aligned with the principal axes of the $K$ tensor, similarly to other groundwater flow codes like MODFLOW (Harbaugh, 2005) and ZOOMQ3D (Jackson and Spink, 2004), their diagonal components are not considered. The code solves the implicit form of the numerical equations in an iterative approach.

Conceptually, this model represents a domain consisting of a set of hydrogeological units that are stacked above each other. Numerically, the abstraction borehole occupies the center of the grid and grid nodes are distributed along lines radiating from the center towards the cylindrical boundary of the domain. Each line represents one radial direction along which the grid spacing increases in a logarithmic pattern from the center to the outer boundary to

provide a precise representation of the radial flow field particularly around the abstraction borehole. The set of nodes situated along different radial directions within one plane represents one hydrological unit. To represent the third dimension, the same distribution of nodes is repeated on different planes a number of times equal to the number of the remaining hydrogeological units. The domain discretization is based, therefore, on a layered cylindrical grid. A fully implicit numerical solution of Equation (1) is calculated for all the grid nodes using a successive overrelaxation scheme. Groundwater flow can be simulated under either confined or unconfined conditions. When unconfined conditions are assumed, the mathematical representation of the system becomes complex due to the non-linearity associated with the movement of the water table (Neuman, 1972; Todsen, 1971). However, this non-linearity is addressed by ignoring the high power terms of the equation representing the movement of the water table as suggested by Rushton and Redshaw (1979). The movement of the water table is hence simulated by introducing an additional set of numerical nodes at the top of the upper layer (Bennett et al., 1990), and by assigning an allocated storage coefficient ($S = S_s b + S_y$) equivalent to the specific yield $S_y$ of the layer. Hydraulically, these nodes are only connected to the nodes of the layer below, which is assumed as confined.

To simulate the transient radial flow field generated by the pumping test, the implemented COOMPuTe model considers a cylindrical domain of radius equal to 10,000 m and thickness of 80 m, centered on the abstraction borehole EA. This large domain was chosen to mitigate the effect of the boundary conditions on the simulated radial flow field around the pumping borehole. Radially, the domain is discretized into 306 nodes. The spacing between nodes increases logarithmically with the distance $r$ from the abstraction borehole from a minimum of 0.1 m up to a maximum of 1840 m. Vertically, the domain covers the saturated thickness of the aquifer from the static water table depth (around 20 m) down to a depth of 100

312 m. This thickness is discretized in four layers whose boundaries were chosen to be consistent
313 with the four flow horizons identified from the flowmeter data (Figure 3). However, different
314 conceptual models of aquifer heterogeneity were considered to estimate characteristic *K* values
315 to each layer as it will be described in the next section.

316 The numerical model is subject to the following set of initial and boundary conditions.
317 Neumann boundary conditions with prescribed flux equal to zero (no flow) are applied to the
318 lateral surface of the cylindrical domain and to its bottom circular. By imposing these
319 conditions, the recharge rate during the duration pumping test is assumed negligible, and it is
320 also assumed that there is no nearby source of groundwater that may affect the responses to
321 pumping in the aquifer. Specified flux, simulating constant groundwater abstraction at a rate
322 similar to the average rate imposed in field conditions (5770 $m^3$/day) was applied to a node at
323 the center of the domain. This node has an area equal to the area of the pumped borehole
324 (diameter equal to 0.73 m) and an assigned storage coefficient equal to one to represent
325 borehole storage effects. This central node is connected with high conductance values to the
326 aquifer nodes located above the maximum depth of the abstraction borehole to take into
327 account the effects of partial aquifer penetration. These values ensure that all connected nodes
328 have the same calculated heads representing the water level inside the pumped borehole. The
329 specified flux was kept active for a simulation time consistent with the duration of the
330 abstraction phase, then switched to zero for the additional simulation time representing the
331 recovery phase. A very small initial time step ($10^{-5}$ seconds) was used at the beginning of each
332 phase and increased logarithmically until the end of the simulation time. Because of the very
333 low hydraulic gradient in the studied area, constant uniform initial heads values were assigned
334 as initial conditions to the nodes of the domain. For simplicity, the effect of borehole losses

due to pumping in the observed and simulated drawdowns of the abstraction borehole are neglected, and radial symmetry is also assumed by imposing $K_r = K_\theta$.

### 3.2.2 Conceptual models and calibration strategies

The numerical model was implemented on the basis of four conceptual models whose main properties are summarized in Table 3. These models consider different aquifer structures (homogenous or layered) as well as different assumptions regarding the isotropy of the *K* field. For each case, input values for the hydrogeological properties (i.e. *K* and *S*) were automatically adjusted with the non-linear parameter estimation software PEST (Doherty, 2015) in order to minimize the objective function:

$$\Phi = \sum_i \left(D_{obs,i} - D_{mod,i}\right)^2 \quad (2)$$

where $D_{obs,i}$ is a measured drawdown and $D_{mod,i}$ is the corresponding simulated value. For the three piezometers, the modelled drawdown ($D_{mod,i}$) is a drawdown simulated at one node of the numerical grid while for the three open boreholes, $D_{mod,i}$ has been obtained by averaging the head values simulated at the grid nodes, located along one vertical gridline, that are in contact, or representing, the open borehole. The total number of measurements used to calculate the objective function for each simulation is in excess of 8,000. Alternative objective functions could be used for the solution of the inverse problem. However, Equation 1 was considered in this work since we focus on the identification of effective hydrogeological parameters with a relatively large support scale. To this end, the sum of squared residuals is an appropriate choice since it tends to put more emphasis on the late time data (i.e. larger drawdown values). Equation 1 is also the default objective function in PEST. Therefore, the optimization strategies and the results of study can be readily transferred to other datasets.

For each model, different calibration strategies were adopted. These differ with respect to the total number of variables, as well as to the number and type (i.e. spatial, numerical or

359 both) of constraints considered in the minimization of Equation 2 (Table 3). Spatial constraints
360 to the vertical variability of $K$ are applied by imposing a structure in the numerical model
361 consisting of four layers each having uniform $K$ values and thicknesses that correspond to those
362 of the flow horizons identified from the flowmeter data (Figure 3. In one model (M4), these
363 deterministic spatial constraints are coupled to numerical constraints in the form of bounds
364 within which PEST can search for the solution of the minimization problem. These bounds are
365 defined by prior information from the interpretation of the packer tests. Details for each model
366 in Table 3 are as follows. Model M1 assumes a homogeneous and isotropic unconfined aquifer
367 and therefore a constrain was imposed in the calibration process such that the same $K$ value is
368 assigned to each layer of the model. Accordingly, the simulated transient radial flow field and
369 simulated drawdowns are controlled only by three input variables: a single $K$ value, the specific
370 storage and the specific yield of the aquifer. Model M2 assumes a 4-layer structure, but
371 isotropic conditions are assumed for the principal components of the hydraulic conductivity
372 tensor in each layer (i.e. $K_r = K_z$). This model also assumes that the storage coefficient is the
373 same in all the layers. Accordingly, a total of six input variables were considered for automatic
374 calibration with PEST, which is conducted under unconstrained conditions at least in terms of
375 range of variability assumed by the variable values. A similar calibration approach is applied
376 for the estimation of the optimal hydrogeological parameters in model M3. However, this
377 model also assumes anisotropy for $K$ in the four layers. Hence, a total of 10 variables are varied
378 to fit the observed drawdowns and minimize Equation 1. Finally, model M4 assumes a layered
379 and anisotropic aquifer similar to M3, but the calibration of $K_r$ and $K_z$ values for each layer
380 takes also into account numerical bounds corresponding to the range of horizontal $K$
381 measurements from the packer test. In this model, multi-scale prior information about the
382 vertical $K$ variability is fully integrated in the calibration of the numerical simulation. In

383 practice, this integration is possible through an upscaling of the local scale prior information,
384 which is performed by setting both spatial (i.e., boundaries between layers) and numerical
385 constraints (i.e., ranges of $K$ values) to the optimal $K$ values in the numerical model.

386 Another approach to include prior information in the calibration process is to add a
387 regularization term directly to the objective function. In the case of Tikhonov regularization,
388 for instance, this term includes a set of constraints on parameter values that need to be
389 formulated to express the expert knowledge that is relevant to a particular problem (Doherty,
390 2015). Mathematical regularization methods included in PEST such as the Tikhonov
391 regularization or singular value decomposition (SVD) allow to address the issue of non-
392 uniqueness in the solution of the inverse problem (Tarantola, 2005), although numerical
393 stability cannot always guaranteed (Doherty, 2015). The approach used in this work to calibrate
394 the different radial flow models also allow to achieve uniqueness in the solution of the inverse
395 problem by providing prior knowledge in the form of constraints for the calibration parameters.
396 However, differently from purely mathematical regularization methods, this knowledge is
397 applied here also at the conceptual level through the imposition of a deterministic layered
398 structure of the aquifer justified by prior experimental evidence. In situations where this
399 evidence is not available or uncertain, a geostatistical inverse method (Kitanidis, 1995; Li et
400 al., 2005; Zimmerman et al., 1998) may be a preferred option since these methods require a
401 minimum amount of prior information regarding the spatial distribution of the parameters.
402

403 **4 Results and discussion**

404 **4.1 Impact of calibration strategies on simulated time-drawdown curves**

405 Examples of comparisons between observed and simulated time-drawdown curves for
406 the abstraction borehole (EA), the open borehole A and the two piezometers in borehole D are

407 presented in Figures 5 – 8 for all the implemented models. Minimum values of the objective
408 function and corresponding calibrated values of the input parameters are also reported in
409 Table 4. Model M1 provides reasonably accurate simulations of the experimental data in some
410 of the monitoring boreholes particularly for the drawdowns measured during the recovery
411 phase of the pumping test. As expected, the assumption of a homogeneous and isotropic $K$
412 distribution seems to be more effective for matching the slopes of the time-drawdown curves
413 at late rather than the early times. The calibrated $K$ is equal to 14 m/day, which corresponds to
414 a transmissivity value of 1120 $m^2$/day. This value is similar to those determined using the Theis
415 analytical model for the interpretation of the late time drawdown data (Table 2). The calculated
416 minimum value for the objective function $\Phi$ is 910 $m^2$ corresponding to a root mean squared
417 error ($RMSE = \sqrt{\Phi/n}$ where $n$ is the total number of observations) of about 0.33 m.
418 Significant errors in the simulated time-drawdown curve are observed at the abstraction
419 borehole (Figure 5a). In particular, observed drawdowns are overestimated for the abstraction
420 phase of the test, while the recovery data is overestimated at early times and underestimated at
421 later times.

422 Results from model M2 (Figure 6) indicate an improvement in the goodness-of-fit of
423 the simulated drawdowns. The objective function value of this model is in fact about 30%
424 lower than the minimum value calculated from the model M1 (639 $m^2$ vs. 910 $m^2$) and the
425 RMSE is 0.28 m. In particular, most of the improvement is observed for the abstraction
426 borehole. This result suggests that the errors observed in model M1 are most probably caused
427 by the oversimplification of the aquifer structure into a single homogeneous layer rather than
428 being systematic errors related to the effects of local heterogeneities and/or head losses not
429 taken into account in the radial model. The consistency of the results when the abstraction

borehole data is not considered for model calibration (see next section) confirms this conclusion. Moreover, because of the very large number of available data for model calibration (>8000 drawdown values), the small increment in the number of parameters from model M1 to model M2 is vastly justified by the data, and overfitting issues are irrelevant in our analysis. Therefore, the improvement in accuracy between the two models indicates that the integration of medium scale deterministic prior information regarding the location of boundaries between layers with different conductivity results in a better characterization of the actual aquifer heterogeneity.

Similar considerations regarding the impact of increasing the complexity of the aquifer structure and the number of calibration parameters can be made for model M3 with respect to M2, as well as for model M3 with respect to M4. Model M3 simulates the abstraction and recovery phases with good accuracy in all the monitored boreholes including the abstraction borehole (Figure 7). The calibrated values for this model correspond to an objective function value that is about 45% lower than the value for model M2 (348 $m^2$ vs. 639 $m^2$), which indicates that the anisotropic $K$ distribution in the layers provides a better representation of the actual aquifer heterogeneity. However, some of the calibrated $K_r$ and $K_z$ values are not consistent with the range of the $K$ values in the packer test data. In particular, calibrated values for $K_z$ in layers 3 and 4 are very high compared to the correspondent values of the component $K_r$ (Table 4). This is an interesting result from our case study, because it highlights one of the major difficulties associated with automatic model calibration due to the non-uniqueness of the solution of the inverse problem in hydrogeological applications. This difficulty is the fact that the sets of identified optimal parameter values may be unrealistic and inconsistent with the hydrogeological setting of the aquifer. To address this issue, prior information needs to be included (e.g. Carrera et al., 2005; Raghavan, 2004; Zhou et al., 2014). This is confirmed by

454 our analysis and in particular by the results of model M4, which, as shown by Figure 8, provides
455 the best fit to the experimental data of all the implemented models ($\Phi$ = 286 m$^2$ corresponding
456 to RMSE = 0.19 m). This results show that the upscaled local $K$ variability can be used
457 effectively as a numerical constraint to improve both the accuracy and the representativeness
458 of the input parameters of a groundwater flow model.

459 A very informative output from PEST is the so-called composite parameter
460 sensitivities. For each parameter, these represent the magnitude of the corresponding column
461 of the Jacobian matrix normalized with respect to the number of observations (Doherty, 2015).
462 The Jacobian matrix contains the values of the derivatives of the observations with respect to
463 the parameters and it is fundamental for the algorithm used for parameter estimation (i.e. the
464 Levenberg-Marquardt method). Relative sensitivities can be calculated by multiplying the
465 composite sensitivities by their corresponding magnitudes of the parameters. In contrast to
466 composite sensitivities, relative values can be used for ranking the sensitivity of different
467 parameters. Calculated relative sensitivity values for the input parameters of the implemented
468 models are within the range $3\times10^{-5}$ – 0.022. Comparisons between values for each model
469 (Figure 9) indicate that the simulated drawdowns are significantly more sensitive to variations
470 in $K$ than in storage. As expected, the aquifer being unconfined, $S_y$ has more impact than $S_s$ in
471 all the models except for model M3 for which comparable sensitivities were estimated for the
472 two parameters (Figure 9a). For the models considering a heterogeneous $K$ distribution in the
473 aquifer (i.e. M2, M3, and M4), simulated drawdown curves are generally more sensitive to the
474 $K$ values assigned to the shallowest layer (layer 1) and to layer 3, while the relative sensitivity
475 of the conductivity of the deepest layer (layer 4) is the lowest except for model M4 (Figure 9b).
476 This result is likely related to the fact that specified fluxes simulating groundwater abstraction
477 are imposed only in layers 1 – 3 of the model. The comparison between relative sensitivity

478 values of $K_r$ and $K_z$ for the two anisotropic models M3 and M4 (Figure 9c) indicates that
479 imposing constraints in the variability of $K$ in the different layers (i.e. as in model M4) affects
480 the sensitivity of the simulated drawdowns. In particular, results indicate that model M4 is
481 significantly most sensitive to the $K_r$ value assigned to the most conductive layer (layer 1).
482 Moreover, the application of constraints based on prior local $K$ information from the packer
483 test results in a homogenization of the relative sensitivities in some of the layers, particularly
484 in layer 4.

485 **4.2 Validation of model M4**

486 Three additional simulations were undertaken to validate model M4 and investigate the
487 impact of the amount of data used for model calibration on the accuracy of simulated drawdown
488 and calibrated hydrogeological values. In the first simulation (M4_noEA in Table 4), model
489 M4 was calibrated without considering the data from borehole EA to investigate the impact of
490 the heterogeneity around the pumping well and unaccounted for well losses on model results.
491 Calibrated storage parameters for this model are very similar to those estimated for the original
492 model M4, as well as the values of $K_r$ for layers 1, 2 and 4 (Table 4). Larger differences are
493 observed for the calibrated values of $K_z$ particularly for layers 2, 3, and 4. However, these
494 discrepancies are not relevant due to the low sensitivity of the model outputs with respect to $K_z$
495 (see also Figure 9c). For instance, the value of 61 m/d calculated for the $K_z$ of layer 2 of model
496 M4-noEA has practically no impact on the simulated drawdown values since the relative
497 sensitivity is equal to $3.1\times10^{-6}$. This value is even lower than the corresponding relative
498 sensitivity estimated for model M4 ($2.7\times10^{-4}$ in Figure 9c). The irrelevancy of the calibrated
499 $K_z$ values on the outputs of model M4-noEA was confirmed by the consistency in the simulated
500 drawdowns when this model was run with an imposed $K_z$ equal to 2 m/d for layer 2 (i.e. the
501 same as M4). A similar test was conducted for the other additional simulations in Table 4 (M4-

502 noBS and M4-noADE), and the results indicated that they are also not sensitive to variations
503 in $K_z$.

504 An important result of model M4-noEA is that drawdown values are simulated with
505 comparable accuracy with respect to model M4 at all the observation points, while a decrease
506 in accuracy is observed for the prediction of the abstraction and the early-time recovery phases
507 at the abstraction borehole (Figure 10a). A sensitivity analysis of all the parameters showed
508 that this discrepancy is caused by the difference in $K_r$ values of layer 3 between the two models
509 (Table 4). In particular, the $K_r$ estimated for layer 3 (0.2 m/d) is an order of magnitude lower
510 than the value in model M4 (2 m/d) and very close to the lower bound imposed on the $K$ values.
511 Although we cannot exclude some influence from unaccounted for well losses, this result
512 suggests that the pumping borehole data provides necessary information to better constrain the
513 vertical $K$ variability.

514 Another simulation (M4-noBC) was performed on the basis of model M4, but with the
515 exclusion of borehole B and piezometers C1 and C2 from the sources of data for automatic
516 calibration. These observation points are the furthest from the pumped borehole, and therefore
517 with this additional model we tested the effect of considering data reflecting larger scale
518 heterogeneities in the pumping test analysis. Simulated drawdowns match the experimental
519 data at the boreholes and piezometers included in the calibration process. In addition, the
520 accuracy of the predictions of the observed time drawdown curves at the excluded locations is
521 comparable to the accuracy of the simulations using all the data (Figure 10b), and a similarity
522 between corresponding sensitive parameters for models M4 and M4-noBC is observed (Table
523 4). This result strengthen the validity of the layered structure and ranges of $K$ values derived
524 from prior information, which appear to be stationary within the scale of investigation of the
525 pumping test.

526 The reliability of the model M4 was further confirmed by the results of another
527 simulation (M4-noADE). This time, the data at boreholes A and E and piezometers D1 and D2
528 in borehole D were excluded from the calibration process. As for the other model M4-noBC,
529 predictions of the drawdown values at the excluded locations (e.g. Figure 10c) are generally as
530 accurate as the simulations from model M4. Calibrated hydraulic parameter values are also
531 very similar to those based on the complete dataset.
532
533 **5 Conclusions**
534 A radial flow numerical model was used to simulate experimental data collected during
535 a pumping test in a dual-porosity unconfined aquifer (the Chalk aquifer in southern England).
536 Different conceptualizations of the aquifer heterogeneity and automatic calibration approaches
537 were tested to evaluate the effect of model complexity and integration of multi-scale
538 hydrogeological data on the accuracy and sensitivity of the simulated responses. Based on the
539 results of this study, the following conclusions can be drawn.
540 The assumption of homogeneity and isotropy in the $K$ distribution (i.e. model M1)
541 resulted in a reasonably accurate simulation of the drawdowns only for certain observation
542 boreholes and/or only for segments of the observed time-drawdown curves. This means that it
543 is not possible to find a unique representative $K$ value for the considered aquifer, as it is
544 assumed by analytical solutions used generally applied for pumping test data interpretation.
545 Medium scale vertical $K$ variations based on flowmeter measurements (model M2)
546 provided prior information for the definition of a deterministic layered aquifer structure that
547 significantly improved the goodness-of-fit of the simulated drawdowns. The broader meaning
548 for this result is that a better conceptualization of the aquifer can be achieved with little extra
549 data that are rather inexpensive in terms of associated costs and time involved.

When the different principal components of the *K* tensor were considered (model M3), a 45% increment in accuracy was obtained. Although in general the responses of the implemented models are more sensitive to variations of the radial component of *K*, this result highlights the importance of estimating also the vertical component for the characterization of dual-porosity unconfined aquifers.

Automatic model calibration can result in unrealistic calibrated values for the hydrogeological parameters. As shown by previous studies, prior information can be effective to address this issue.

A simple upscaling approach was applied to integrate small-scale *K* data based on packer testing in the automatic calibration process (model M4). Providing realistic bounds to the variability of the *K* values in the model layers resulted in a further significant improvement in accuracy of the simulated drawdown. The predictive ability of this model was also tested to validate the reliability of the conceptual model derived from the multi-scale prior information.

This work provides general insights for the interpretation of pumping tests in heterogeneous and hydraulically complex aquifers for which the assumptions of the traditional methods of interpretation based on analytical solutions do not hold. The results suggest that the most representative hydrogeological characterization of these aquifers can be achieved with the integration of multi-scale data.

**Acknowledgments.** The authors would like to thank the reviewers for their valuable comments to improve the manuscript. This paper is published by permission of the executive director of the British Geological Survey.

573 **References**

574 Allen, D.J., Brewerton, L.J., Coleby, L.M., Gibbs, B.R., Lewis, M.A., MacDonald, A.M.,
575 Wagstaff, S.J., Williams, A.T., 1997. The physical properties of major aquifers in
576 England and Wales (BGS Technical Report No. WD/97/34). British Geological
577 Survey.
578 Barrash, W., Dougherty, M.E., 1997. Modeling Axially Symmetric and Nonsymmetric Flow
579 to a Well with MODFLOW, and Application to Goddard2 Well Test, Boise, Idaho.
580 Ground Water 35, 602–611. https://doi.org/10.1111/j.1745-6584.1997.tb00125.x
581 Bear, J., 2007. Hydraulics of Groundwater. Dover Publications, Mineola, NY.
582 Beckie, R., Harvey, C.F., 2002. What does a slug test measure: An investigation of
583 instrument response and the effects of heterogeneity. Water Resour. Res. 38, 1290.
584 https://doi.org/10.1029/2001WR001072
585 Bennett, G.D., Reilly, T.E., Hill, M.C., 1990. Technical training notes in ground-water
586 hydrology: Radial flow to a well. US Department of the Interior, US Geological
587 Survey.
588 Bianchi, M., 2017. Validity of flowmeter data in heterogeneous alluvial aquifers. Advances
589 in Water Resources 102, 29–44. https://doi.org/10.1016/j.advwatres.2017.01.003
590 Bloomfield, J., 1996. Characterisation of hydrogeologically significant fracture distributions
591 in the Chalk: an example from the Upper Chalk of southern England. Journal of
592 Hydrology 184, 355–379. https://doi.org/10.1016/0022-1694(95)02954-0
593 Bohling, G.C., Liu, G., Dietrich, P., Butler, J.J., 2016. Reassessing the MADE direct-push
594 hydraulic conductivity data using a revised calibration procedure. Water Resour. Res.
595 52, 8970–8985. https://doi.org/10.1002/2016WR019008

596 Butler, A.P., Mathias, S.A., Gallagher, A.J., Peach, D.W., Williams, A.T., 2009. Analysis of
597 flow processes in fractured chalk under pumped and ambient conditions (UK).
598 Hydrogeol J 17, 1849–1858. https://doi.org/10.1007/s10040-009-0477-4
599 Carrera, J., Alcolea, A., Medina, A., Hidalgo, J., Slooten, L.J., 2005. Inverse problem in
600 hydrogeology. Hydrogeol J 13, 206–222. https://doi.org/10.1007/s10040-004-0404-7
601 Chen, C., Jiao, J.J., 1999. Numerical Simulation of Pumping Tests in Multilayer Wells with
602 Non-Darcian Flow in the Wellbore. Ground Water 37, 465–474.
603 https://doi.org/10.1111/j.1745-6584.1999.tb01126.x
604 Cheng, C., Chen, X., 2007. Evaluation of methods for determination of hydraulic properties
605 in an aquifer–aquitard system hydrologically connected to a river. Hydrogeology
606 Journal 15, 669–678. https://doi.org/10.1007/s10040-006-0135-z
607 Cooper, H.H., Jacob, C.E., 1946. A generalized graphical method for evaluating formation
608 constants and summarizing well-field history. Transactions, American Geophysical
609 Union 27, 526. https://doi.org/10.1029/TR027i004p00526
610 Doherty, J., 2015. Calibration and Uncertainty Analysis for Complex Environmental Models.
611 Watermark Numerical Computing, Brisbane, Australia.
612 Domenico, P.A., Schwartz, F.W., 1997. Physical and Chemical Hydrogeology, 2 edition. ed.
613 Wiley, New York, NY.
614 Fetter, C.W., 2000. Applied Hydrogeology, 4 edition. ed. Pearson, Upper Saddle River, NJ.
615 Fogg, G.E., Noyes, C.D., Carle, S.F., 1998. Geologically based model of heterogeneous
616 hydraulic conductivity in an alluvial setting. Hydrogeology Journal 6, 131–143.
617 https://doi.org/10.1007/s100400050139

618 Gelhar, L.W., Axness, C.L., 1983. Three-dimensional stochastic analysis of macrodispersion
619 in aquifers. Water Resour. Res. 19, 161–180.
620 https://doi.org/10.1029/WR019i001p00161
621 Halford, K.J., Yobbi, D., 2006. Estimating Hydraulic Properties Using a Moving-Model
622 Approach and Multiple Aquifer Tests. Ground Water 44, 284–291.
623 https://doi.org/10.1111/j.1745-6584.2005.00109.x
624 Harbaugh, A.W., 2005. MODFLOW-2005, The U.S. Geological Survey Modular Ground-
625 Water Model—the Ground-Water Flow Process, U.S. Geological Survey Techniques
626 and Methods.
627 Ireson, A.M., Mathias, S.A., Wheater, H.S., Butler, A.P., Finch, J., 2009. A model for flow in
628 the chalk unsaturated zone incorporating progressive weathering. Journal of
629 Hydrology 365, 244–260. https://doi.org/10.1016/j.jhydrol.2008.11.043
630 Jackson, C.R., Spink, A.E.F., 2004. User's manual for the groundwater flow model
631 ZOOMQ3D (No. IR/04/140), British Geological Survey Internal Report. British
632 Geological Survey.
633 Kaleris, V., Hadjitheodorou, C., Demetracopoulos, A.C., 1995. Numerical simulation of field
634 methods for estimating hydraulic conductivity and concentration profiles. Journal of
635 Hydrology 171, 319–353. https://doi.org/10.1016/0022-1694(94)06012-T
636 Kitanidis, P.K., 1995. Quasi-Linear Geostatistical Theory for Inversing. Water Resour. Res.
637 31, 2411–2419. https://doi.org/10.1029/95WR01945
638 Koltermann, C.E., Gorelick, S.M., 1996. Heterogeneity in Sedimentary Deposits: A Review
639 of Structure-Imitating, Process-Imitating, and Descriptive Approaches. Water Resour.
640 Res. 32, 2617–2658. https://doi.org/10.1029/96WR00025

641 Kruseman, G.P., de Ridder, N.A., 1990. Analysis and Evaluation of Pumping Test Data, 2nd
642 ed. International Institute for Land Reclamation and Improvement (ILRI), The
643 Netherlands.
644 Lebbe, L., Mahauden, M., Breuck, W.D., 1992. Execution Of A Triple Pumping Test And
645 Interpretation By An Inverse Numerical Model. HYJO 1, 20–34.
646 https://doi.org/10.1007/PL00010967
647 Lee, S.-Y., Carle, S.F., Fogg, G.E., 2007. Geologic heterogeneity and a comparison of two
648 geostatistical models: Sequential Gaussian and transition probability-based
649 geostatistical simulation. Advances in Water Resources 30, 1914–1932.
650 https://doi.org/10.1016/j.advwatres.2007.03.005
651 Leven, C., Dietrich, P., 2006. What information can we get from pumping tests?-comparing
652 pumping test configurations using sensitivity coefficients. Journal of Hydrology 319,
653 199–215. https://doi.org/10.1016/j.jhydrol.2005.06.030
654 Li, L., Zhou, H., Hendricks Franssen, H.-J., Gómez-Hernández, J.J., 2012. Modeling transient
655 groundwater flow by coupling ensemble Kalman filtering and upscaling. Water
656 Resour. Res. 48, W01537. https://doi.org/10.1029/2010WR010214
657 Li, W., Nowak, W., Cirpka, O.A., 2005. Geostatistical inverse modeling of transient pumping
658 tests using temporal moments of drawdown. Water Resour. Res. 41, W08403.
659 https://doi.org/10.1029/2004WR003874
660 MacDonald, A.M., Allen, D.J., 2001. Aquifer properties of the chalk of England. Quarterly
661 Journal of Engineering Geology and Hydrogeology 34, 371–384.
662 https://doi.org/10.1144/qjegh.34.4.371
663 Mansour, M., Hughes, A.G., Spink, A.E.F., 2007. User manual for the layered R-theta
664 numerical model (Publication - Report No. (OR/07/029). British Geological Survey.

665 Mansour, M.M., Hughes, A.G., Spink, A.E.F., Riches, J., 2011. Pumping test analysis using a
666 layered cylindrical grid numerical model in a complex, heterogeneous chalk aquifer.
667 Journal of Hydrology 401, 14–21. https://doi.org/10.1016/j.jhydrol.2011.02.005
668 Martinez-Landa, L., Carrera, J., 2005. An analysis of hydraulic conductivity scale effects in
669 granite (Full-scale Engineered Barrier Experiment (FEBEX), Grimsel, Switzerland).
670 Water Resour. Res. 41, W03006. https://doi.org/10.1029/2004WR003458
671 Mathias, S.A., Butler, A.P., 2006. Linearized Richards' equation approach to pumping test
672 analysis in compressible aquifers. Water Resour. Res. 42, W06408.
673 https://doi.org/10.1029/2005WR004680
674 Mathias, S.A., Butler, A.P., Peach, D.W., Williams, A.T., 2007. Recovering tracer test input
675 functions from fluid electrical conductivity logging in fractured porous rocks. Water
676 Resour. Res. 43, W07443. https://doi.org/10.1029/2006WR005455
677 Maurice, L.D., Atkinson, T.C., Barker, J.A., Williams, A.T., Gallagher, A.J., 2012. The
678 nature and distribution of flowing features in a weakly karstified porous limestone
679 aquifer. Journal of Hydrology 438–439, 3–15.
680 https://doi.org/10.1016/j.jhydrol.2011.11.050
681 Meier, P.M., Carrera, J., Sánchez-Vila, X., 1998. An evaluation of Jacob's Method for the
682 interpretation of pumping tests in heterogeneous formations. Water Resources
683 Research 34, 1011–1025. https://doi.org/10.1029/98WR00008
684 Moench, A.F., 2003. Estimation of hectare-scale soil-moisture characteristics from aquifer-
685 test data. Journal of Hydrology, Recent Advances in Aquifer Hydraulics and Their
686 Applications to Aquifer and Vadose Zone Characterization, Remediation, and
687 Dewatering 281, 82–95. https://doi.org/10.1016/S0022-1694(03)00202-6

688 Mohamed, A., Rushton, K., 2006. Horizontal wells in shallow aquifers: Field experiment and
689 numerical model. Journal of Hydrology 329, 98–109.
690 https://doi.org/10.1016/j.jhydrol.2006.02.006
691 Nativ, R., Adar, E., Assaf, L., Nygaard, E., 2003. Characterization of the Hydraulic
692 Properties of Fractures in Chalk. Ground Water 41, 532–543.
693 https://doi.org/10.1111/j.1745-6584.2003.tb02387.x
694 Neuman, S.P., 1972. Theory of flow in unconfined aquifers considering delayed response of
695 the water table. Water Resour. Res. 8, 1031–1045.
696 https://doi.org/10.1029/WR008i004p01031
697 Neuman, S.P., Di Federico, V., 2003. Multifaceted nature of hydrogeologic scaling and its
698 interpretation. Rev. Geophys. 41, 1014. https://doi.org/10.1029/2003RG000130
699 Odén, M., Niemi, A., 2006. From well-test data to input to stochastic continuum models:
700 effect of the variable support scale of the hydraulic data. Hydrogeol J 14, 1409–1422.
701 https://doi.org/10.1007/s10040-006-0063-y
702 Oehlmann, S., Geyer, T., Licha, T., Birk, S., 2013. Influence of aquifer heterogeneity on karst
703 hydraulics and catchment delineation employing distributive modeling approaches.
704 Hydrol. Earth Syst. Sci. 17, 4729–4742. https://doi.org/10.5194/hess-17-4729-2013
705 Oliver, D.S., Chen, Y., 2011. Recent progress on reservoir history matching: a review.
706 Comput Geosci 15, 185–221. https://doi.org/10.1007/s10596-010-9194-2
707 Owen, M., Robinson, V.K., 1978. Characteristics and yield in fissured chalk, in: Thames
708 Groundwater Scheme. Thomas Telford Publishing, pp. 33–49.
709 https://doi.org/10.1680/tgs.00605.0003

Poeter, E., Gaylord, D.R., 1990. Influence of Aquifer Heterogeneity on Contaminant Transport at the Hanford Site. Ground Water 28, 900–909. https://doi.org/10.1111/j.1745-6584.1990.tb01726.x

Price, M., 1987. Fluid flow in the Chalk of England. Geological Society, London, Special Publications 34, 141–156.

Price, M., Williams, A., 1993. A pumped double-packer system for use in aquifer evaluation and groundwater sampling. Proceedings of the Institution of Civil Engineers - Water Maritime and Energy 101, 85–92. https://doi.org/10.1680/iwtme.1993.23589

Raghavan, R., 2009. Complex geology and pressure tests. Journal of Petroleum Science and Engineering 69, 181–188. https://doi.org/10.1016/j.petrol.2009.08.010

Raghavan, R., 2006. Some observations on the scale dependence of permeability by pumping tests. Water Resources Research 42, n/a-n/a. https://doi.org/10.1029/2005WR004166

Raghavan, R., 2004. A review of applications to constrain pumping test responses to improve on geological description and uncertainty. Rev. Geophys. 42, RG4001. https://doi.org/10.1029/2003RG000142

Raghavan, R., Dixon, T.N., Thomas, O.O., 2002. Effect of Scale Up and Aggregation on the Use of Well Tests To Identify Geological Properties. Society of Petroleum Engineers. https://doi.org/10.2118/77452-MS

Rovey, C.W., Cherkauer, D.S., 1995. Scale Dependency of Hydraulic Conductivity Measurements. Ground Water 33, 769–780. https://doi.org/10.1111/j.1745-6584.1995.tb00023.x

Rushton, K.R., 2003. Groundwater Hydrology: Conceptual and Computational Models, 1 edition. ed. Wiley-Blackwell, Chichester.

733 Rushton, K.R., Chan, Y.K., 1976. Pumping Test Analysis When Parameters Vary with
734 Depth. Ground Water 14, 82–87. https://doi.org/10.1111/j.1745-6584.1976.tb03637.x
735 Rushton, K.R., Redshaw, S.C., 1979. Seepage and groundwater flow, numerical analysis by
736 analog and digital methods. Wiley, New York.
737 Sanchez-Vila, X., Guadagnini, A., Carrera, J., 2006. Representative hydraulic conductivities
738 in saturated groundwater flow. Reviews of Geophysics 44.
739 https://doi.org/10.1029/2005RG000169
740 Sánchez-Vila, X., Meier, P.M., Carrera, J., 1999. Pumping tests in heterogeneous aquifers:
741 An analytical study of what can be obtained from their interpretation using Jacob's
742 Method. Water Resour. Res. 35, 943–952. https://doi.org/10.1029/1999WR900007
743 Schad, H., Teutsch, G., 1994. Effects of the investigation scale on pumping test results in
744 heterogeneous porous aquifers. Journal of Hydrology 159, 61–77.
745 https://doi.org/10.1016/0022-1694(94)90249-6
746 Schroth, B., Narasimhan, T.N., 1997. Application of a Numerical Model in the Interpretation
747 of a Leaky Aquifer Test. Ground Water 35, 371–375. https://doi.org/10.1111/j.1745-
748 6584.1997.tb00096.x
749 Schulze-Makuch, D., Carlson, D.A., Cherkauer, D.S., Malik, P., 1999. Scale Dependency of
750 Hydraulic Conductivity in Heterogeneous Media. Ground Water 37, 904–919.
751 https://doi.org/10.1111/j.1745-6584.1999.tb01190.x
752 Singh, V.S., 2000. Well storage effect during pumping tests in an aquifer of low permeability.
753 Hydrological Sciences Journal 45, 589–594.
754 https://doi.org/10.1080/02626660009492359
755 Tarantola, A., 2005. Inverse problem theory and methods for model parameter estimation.
756 SIAM, Society for Industrial and Applied Mathematics, Philadelphia, Pa.

Theis, C.V., 1935. The relation between the lowering of the Piezometric surface and the rate and duration of discharge of a well using ground-water storage. Trans. Am. Geophys. Union 16, 519–524. https://doi.org/10.1029/TR016i002p00519

Thorbjarnarson, K. w., Huntley, D., McCarty, J. j., 1998. Absolute Hydraulic Conductivity Estimates from Aquifer Pumping and Tracer Tests in a Stratified Aquifer. Ground Water 36, 87–97. https://doi.org/10.1111/j.1745-6584.1998.tb01068.x

Todsen, M., 1971. On the solution of transient free-surface flow problems in porous media by finite-difference methods. Journal of Hydrology 12, 177–210. https://doi.org/10.1016/0022-1694(71)90005-9

Wheater, H.S., Peach, D., 2004. Developing interdisciplinary science for integrated catchment management: the UK lowland catchment research (LOCAR) programme. International Journal of Water Resources Development 20, 369–385. https://doi.org/10.1080/0790062042000248565

Williams, A., Bloomfield, J., Griffiths, K., Butler, A., 2006. Characterising the vertical variations in hydraulic conductivity within the Chalk aquifer. Journal of Hydrology, Hydro-ecological functioning of the Pang and Lambourn catchments, UKResults from the Lowland Catchment Research (LOCAR) initiative 330, 53–62. https://doi.org/10.1016/j.jhydrol.2006.04.036

Woods, M.A., Newell, A.J., Haslam, R.B., Farrant, A.R., Smith, H., 2015. A physical property model of the Chalk of Southern England [WWW Document]. URL http://nora.nerc.ac.uk/510117/ (accessed 3.9.17).

Yeh, H.-D., Chang, Y.-C., 2013. Recent advances in modeling of well hydraulics. Advances in Water Resources, 35th Year Anniversary Issue 51, 27–51. https://doi.org/10.1016/j.advwatres.2012.03.006

781 Zhou, H., Gómez-Hernández, J.J., Li, L., 2014. Inverse methods in hydrogeology: Evolution
782 and recent trends. Advances in Water Resources 63, 22–37.
783 https://doi.org/10.1016/j.advwatres.2013.10.014
784 Zimmerman, D.A., de Marsily, G., Gotway, C.A., Marietta, M.G., Axness, C.L., Beauheim,
785 R.L., Bras, R.L., Carrera, J., Dagan, G., Davies, P.B., Gallegos, D.P., Galli, A.,
786 Gómez-Hernández, J., Grindrod, P., Gutjahr, A.L., Kitanidis, P.K., Lavenue, A.M.,
787 McLaughlin, D., Neuman, S.P., RamaRao, B.S., Ravenne, C., Rubin, Y., 1998. A
788 comparison of seven geostatistically based inverse approaches to estimate
789 transmissivities for modeling advective transport by groundwater flow. Water Resour.
790 Res. 34, 1373–1413. https://doi.org/10.1029/98WR00003
791
792

793 Table 1. Completion details of the abstraction and observation boreholes. Additional details

794 are provided by Williams et al. (2006).

| Borehole | Borehole diameter [mm] | Piezometer depth [m] | Max depth [m] | Radial distance [m] |
|---|---|---|---|---|
| A | 143 | Open borehole | 100 | 31.9 |
| B | 143 | Open borehole | 100 | 53.9 |
| C | 194 | C1: 40.1 – 39.6 | 40 | 50.1 |
| | | C2: 30.1 – 29.8 | | 49.9 |
| D | 194 | D1: 40.0 – 39.8 | 40 | 38.3 |
| | | D2: 27.0 – 26.8 | | 38.2 |
| E | 143 | Open borehole | 100 | 36.8 |
| F | 194 | F1: 40.0 – 39.8 | 41 | 26.7 |
| | | F2: 30.0 – 29.8 | | 26.6 |
| EA | 760 | Open borehole | 86 | – |

795

796

797

798

799 Table 2. Aquifer properties estimated with the application of the Theis analytical solution to
800 the data collected during the abstraction phase of the pumping test.

| **Borehole** | **Transmissivity [$m^2$/day]** | | **Storage coefficient [-]** | |
|---|---|---|---|---|
| | Early times | Later times | Early times | Later times |
| A | 1400 | 1000 | 0.0015 | 0.004 |
| B | 1800 | 1100 | 0.0006 | 0.0025 |
| E | 1400 | 1000 | 0.0012 | 0.0032 |

801
802
803
804
805
806 Table 3. Conceptual models properties and their calibration strategies.

| **Model** | **Number of parameters considered in PEST** | **Constrains on calibration parameters** | **Aquifer *K*** | **Prior hydrogeological information** |
|---|---|---|---|---|
| M1 | 3 | none | Homogeneous, isotropic | – |
| M2 | 6 | spatial | Heterogeneous, isotropic | Flowmeter |
| M3 | 10 | spatial | Heterogeneous, anisotropic | Flowmeter |
| M4 | 10 | spatial and numerical | Heterogeneous, anisotropic | Flowmeter Packer test |

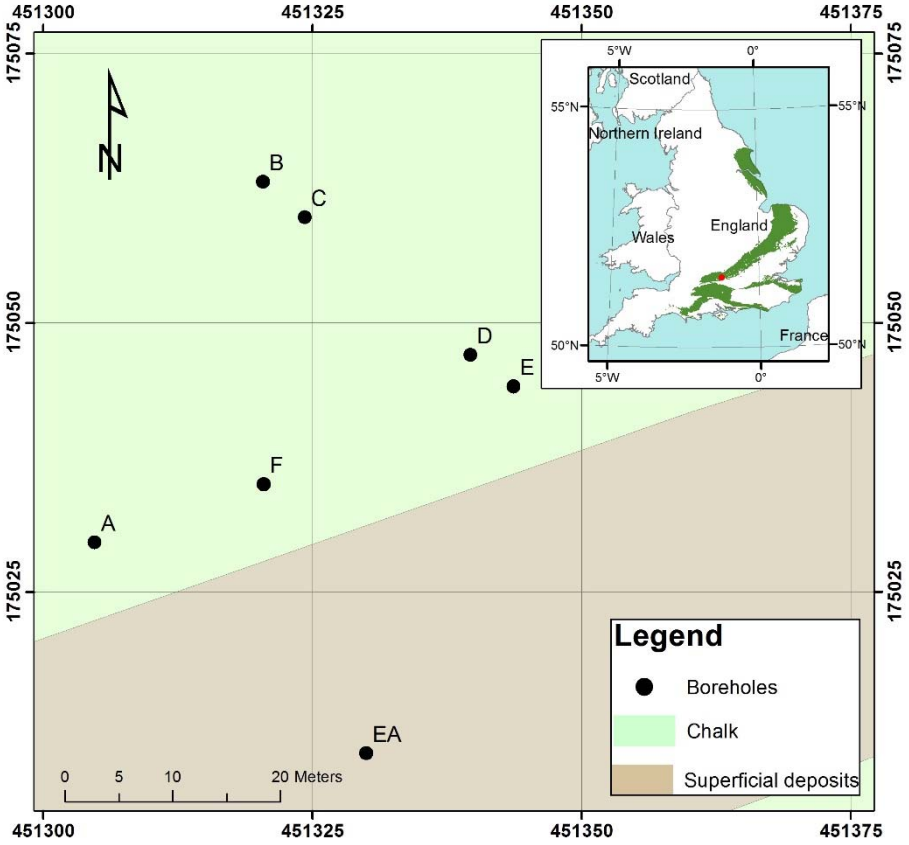
807
808

809 Table 4. Hydrogeological parameter values obtained at the end of the calibration process and
810 corresponding objective function value (Φ).

| **Model** | **$S_s$ [$m^{-1}$ $\times 10^{-6}$]** | **$S_y$** | **Hydraulic conductivity *K* [m/d]** | | | | | | | | **Φ [$m^2$]** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | **Layer 1** | | **Layer 2** | | **Layer 3** | | **Layer 4** | | |
| | | | **$K_r$** | **$K_z$** | **$K_r$** | **$K_z$** | **$K_r$** | **$K_z$** | **$K_r$** | **$K_z$** | |
| M1 | 1.4 | 0.001 | 14 | | | | | | | | 910 |
| M2 | 1.3 | 0.001 | 14 | | 12 | | 23 | | 0.4 | | 639 |
| M3 | 7.7 | 0.0007 | 27 | 2 | 6 | 5 | 19 | 32 | 0.6 | 24 | 348 |
| M4 | 0.1 | 0.001 | 73 | 38 | 1 | 2 | 2 | 10 | 0.1 | 0.2 | 286 |
| M4-noEA | 0.1 | 0.001 | 76 | 38 | 1 | 61 | 0.2 | 0.1 | 0.2 | 0.01 | 233 |
| M4-noBC | 0.1 | 0.001 | 63 | 190 | 1 | 61 | 6 | 3 | 0.1 | 0.2 | 213 |
| M4-noADE | 0.1 | 0.001 | 68 | 38 | 1 | 61 | 5 | 3 | 0.1 | 0.2 | 195 |

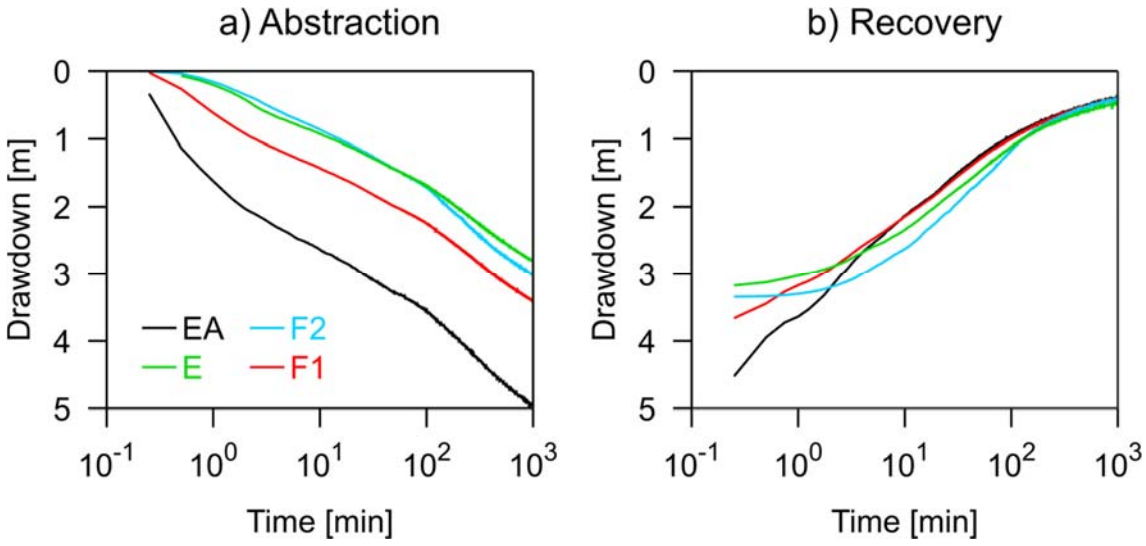
811

812




813

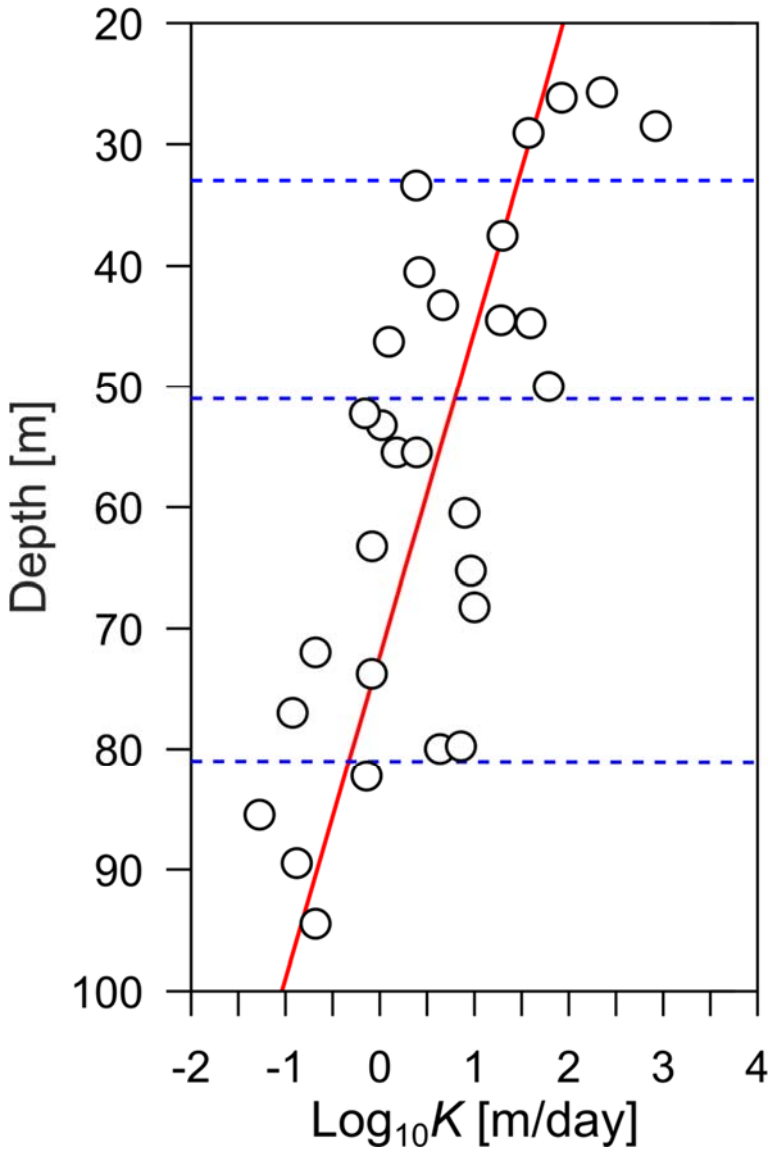
814 Figure 1. Locations of the pumping borehole (EA) and of the six boreholes (A – F) monitored
815 during the pumping test. Coordinates refer to the British National grid (m). The inset figure
816 shows the location of the site relative to the United Kingdom. Chalk outcrops are shown in
817 green.




818

819 Figure 2. Time-drawdown curves for the abstraction borehole (EA), the open borehole E and

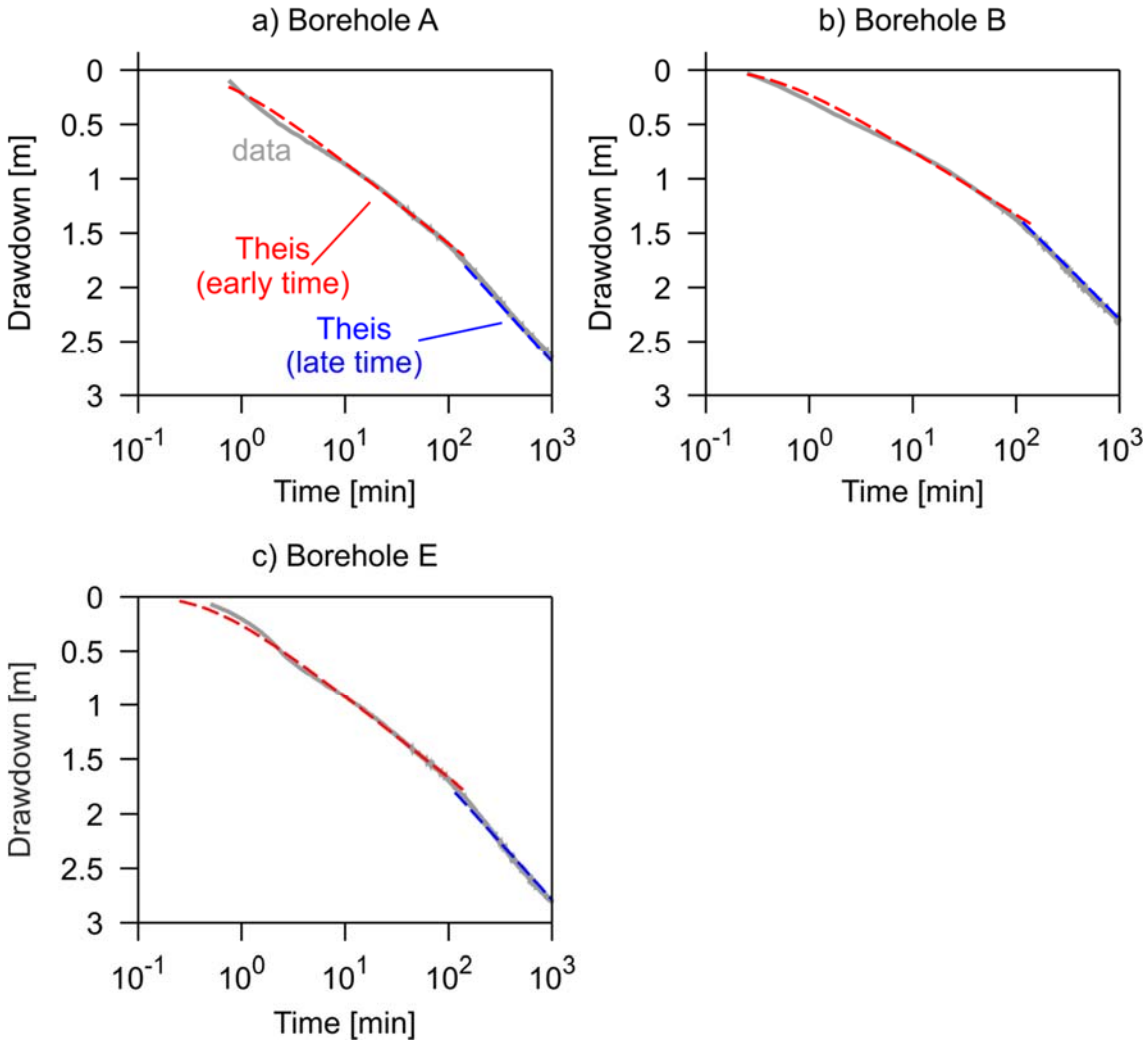
820 for the two piezometers in borehole F during the abstraction (a) and recovery phases (b).




821

822 Figure 3. Vertical hydraulic conductivity profile estimated with the packer test described by
823 Williams et al. (2006). The red line represents the best fitted logarithmic regression model.
824 Blue horizontal lines indicate the boundaries of the four flow horizons identified from the
825 flowmeter log in borehole A.

826




827

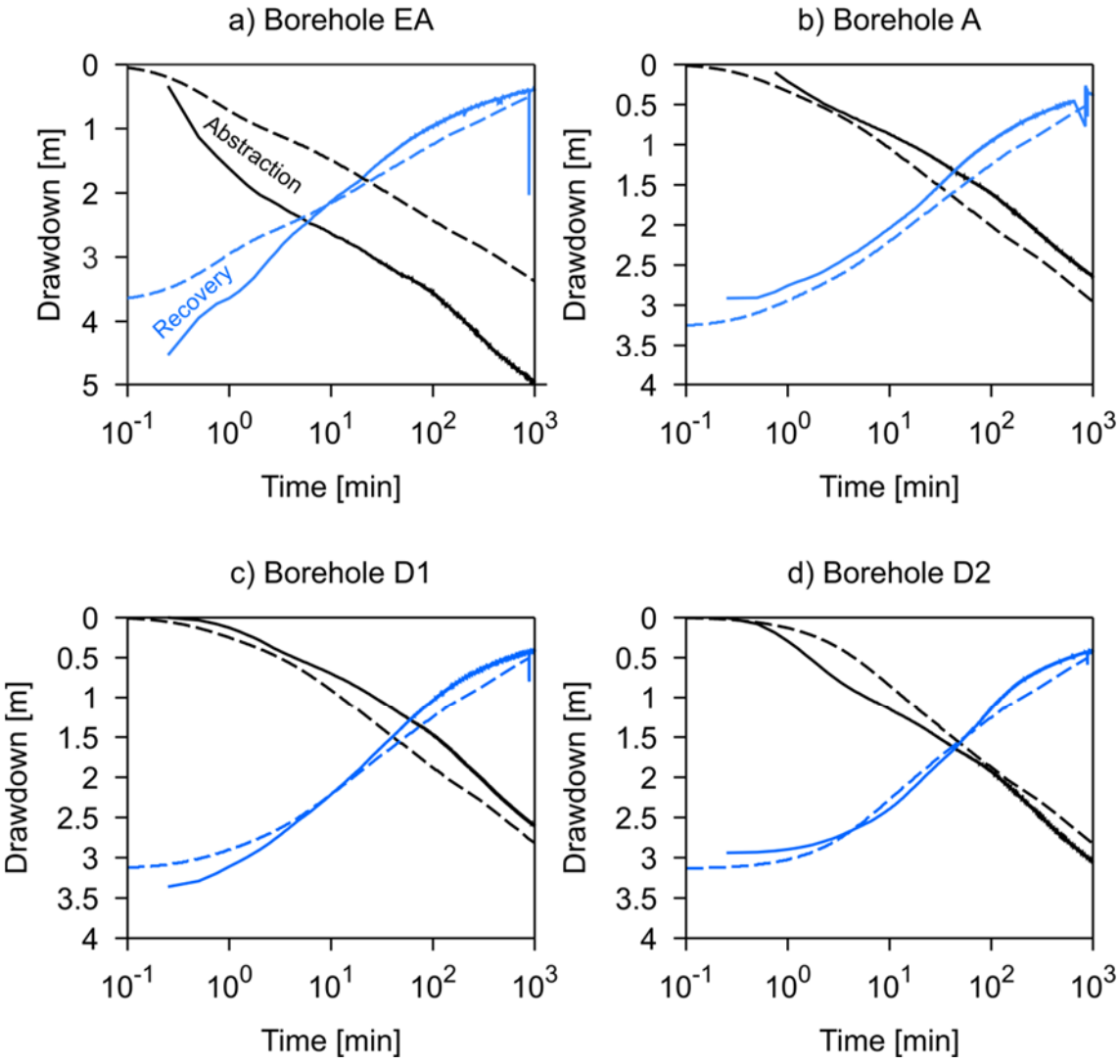
828

829 Figure 4. Examples of interpretation of the pumping test data (solid grey line) according to the

830 Theis analytical solution. For each borehole, two segments corresponding to early (red dashed

831 line) and late (blue dashed line) times were considered.

832



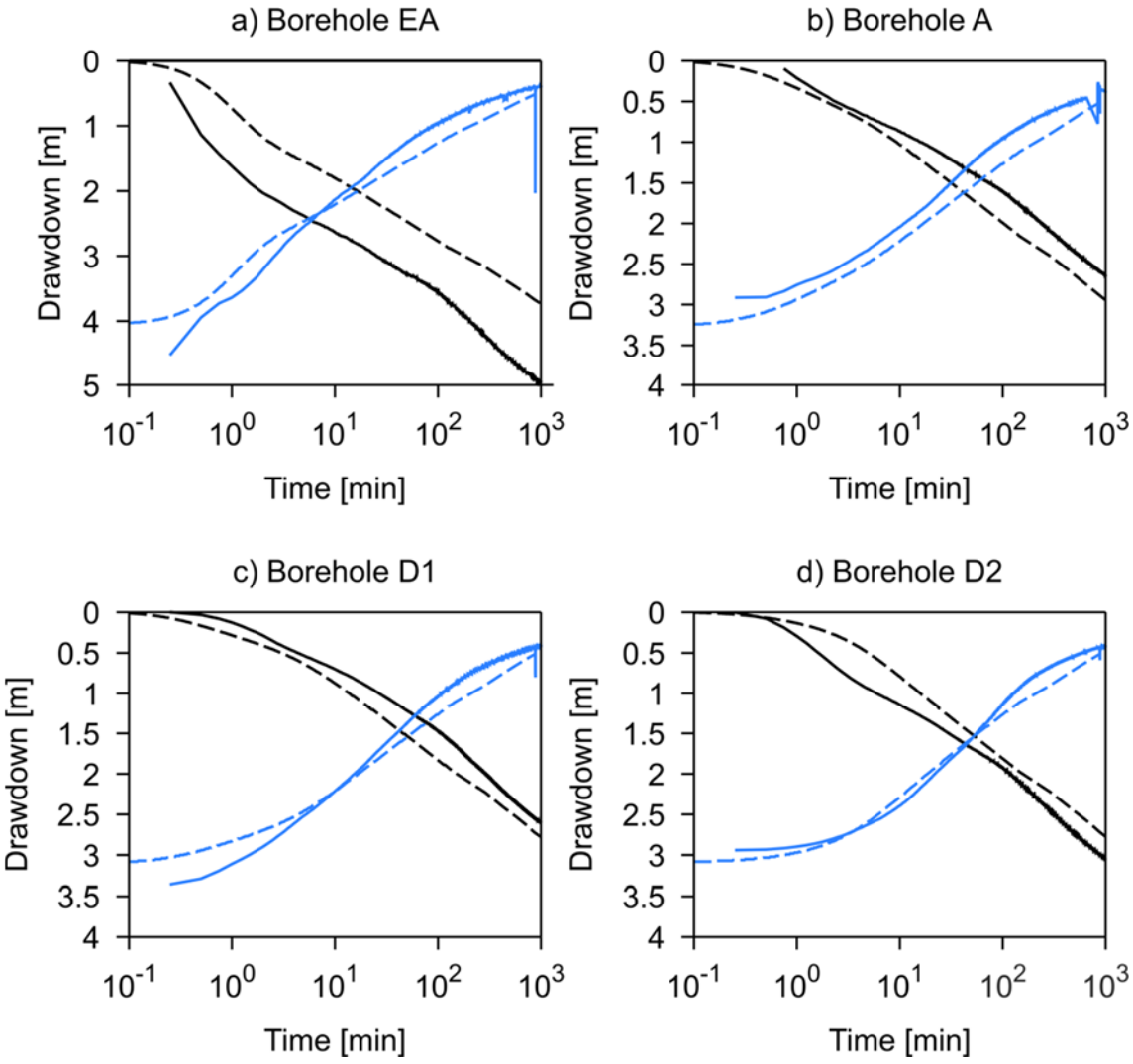

833

834 Figure 5. Model M1. Comparisons between observed (solid lines) and simulated (dashed lines)

835 drawdowns for the abstraction (black lines) and recovery (blue lines) phases.

836

837

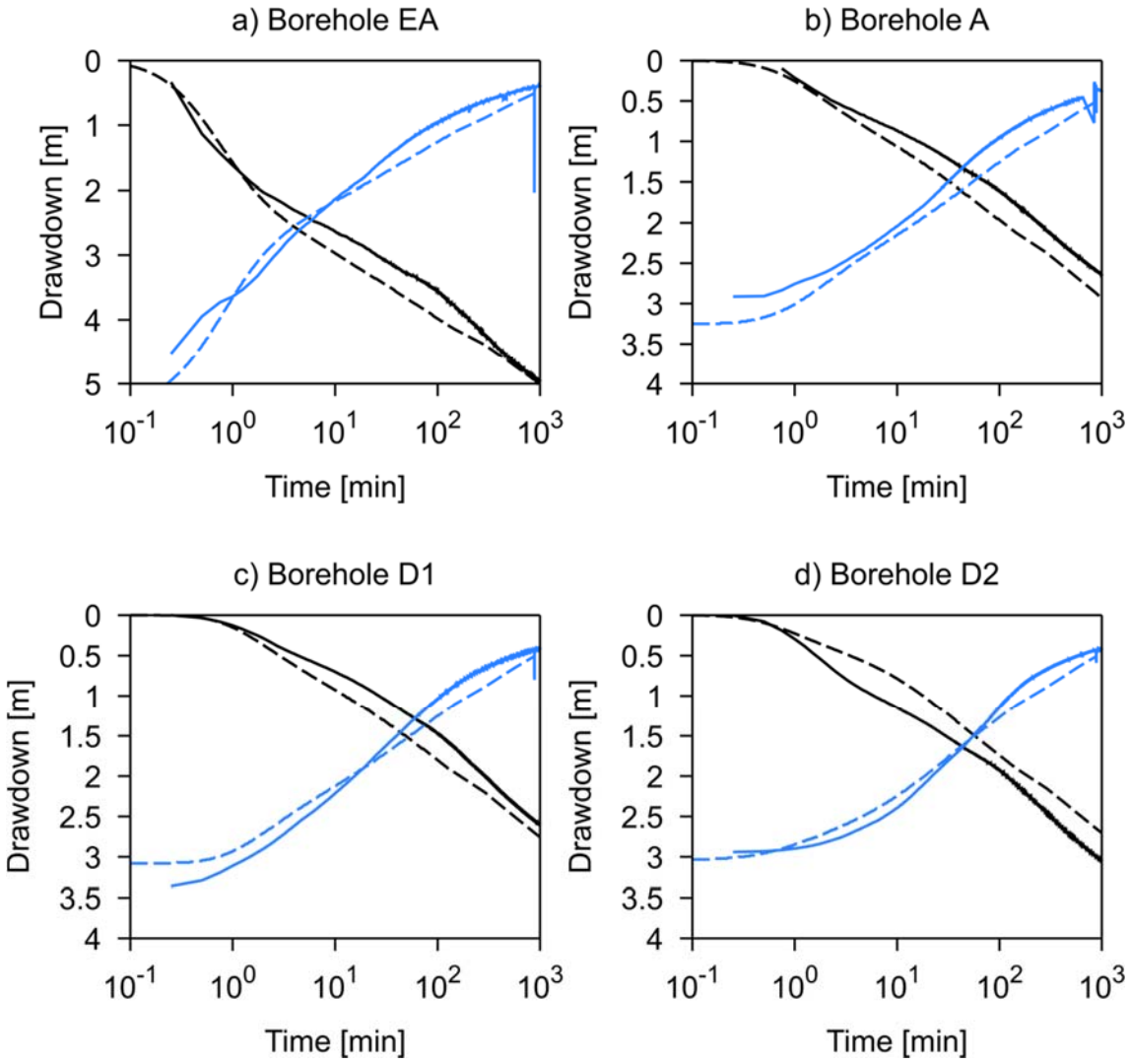



838

839 Figure 6. Model M2. Comparisons between observed (solid lines) and simulated (dashed lines)

840 drawdowns for the abstraction (black lines) and recovery (blue lines) phases.

841

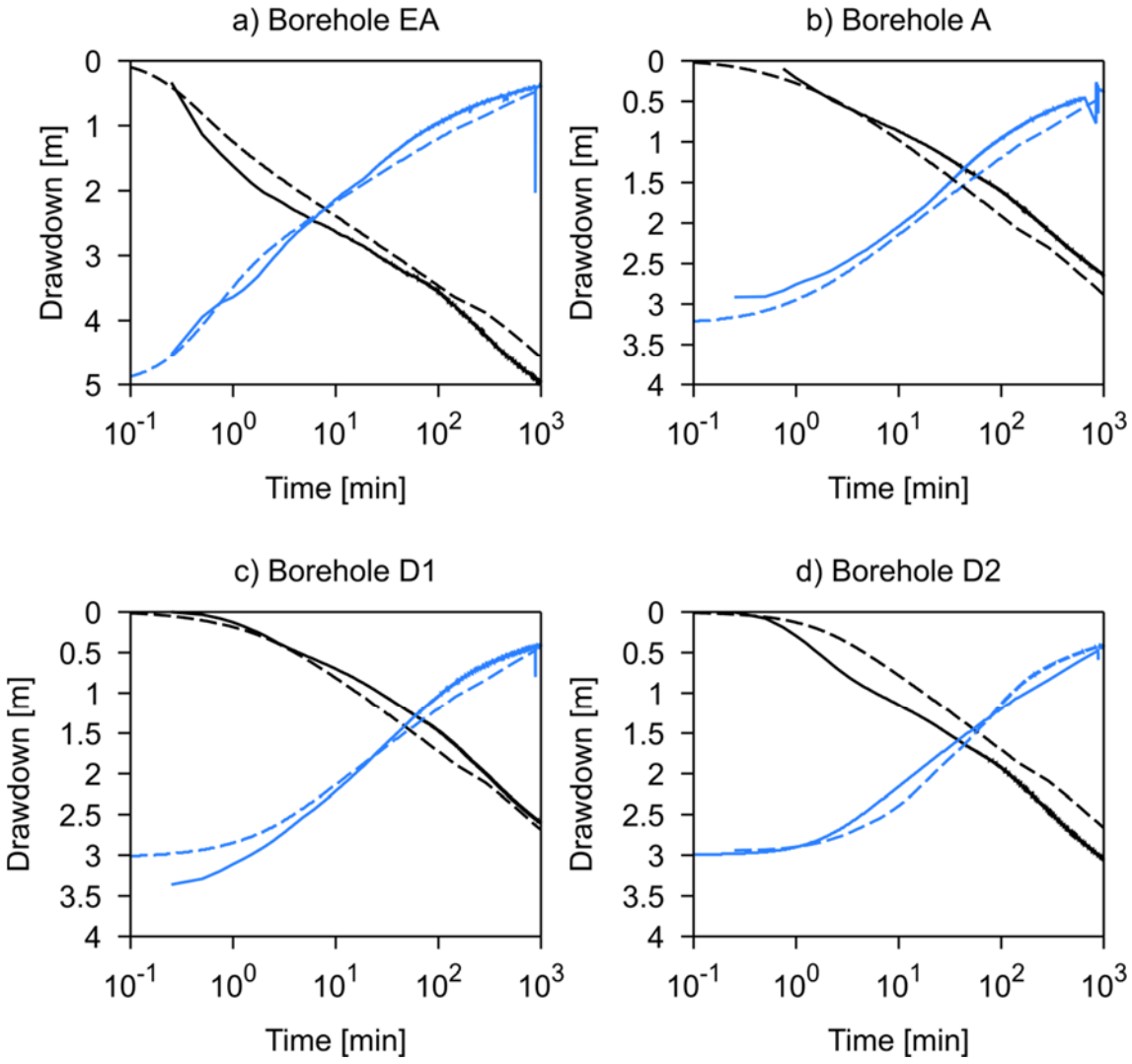



842

843 Figure 7. Model M3. Comparisons between observed (solid lines) and simulated (dashed lines)

844 drawdowns for the abstraction (black lines) and recovery (blue lines) phases.

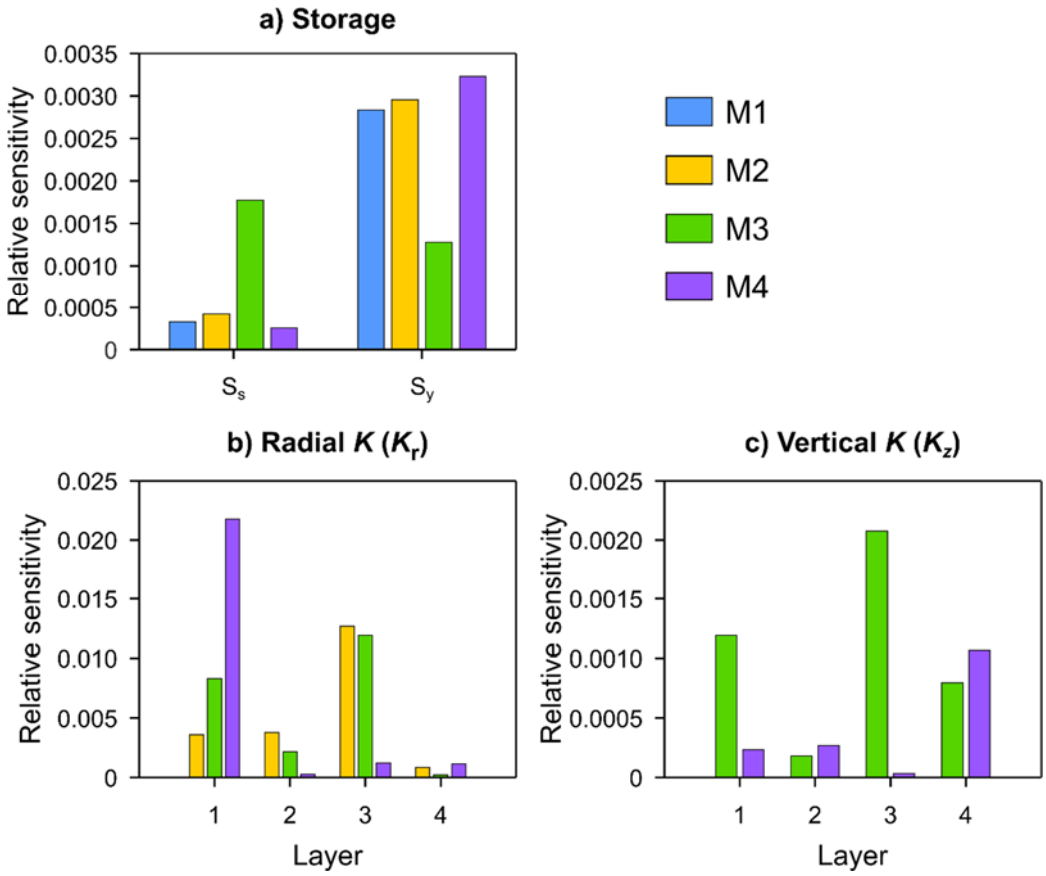
845




846

847 Figure 8. Model M4. Comparisons between observed (solid lines) and simulated (dashed lines)

848 drawdowns for the abstraction (black lines) and recovery (blue lines) phases.




849

850 Figure 9. Relative composite sensitivities of models input parameters: (a) specific storage and
851 the specific yield; (b) radial hydraulic conductivity for models M2, M3 and M4; (c) radial and
852 vertical conductivities for models M3 and M4.

853

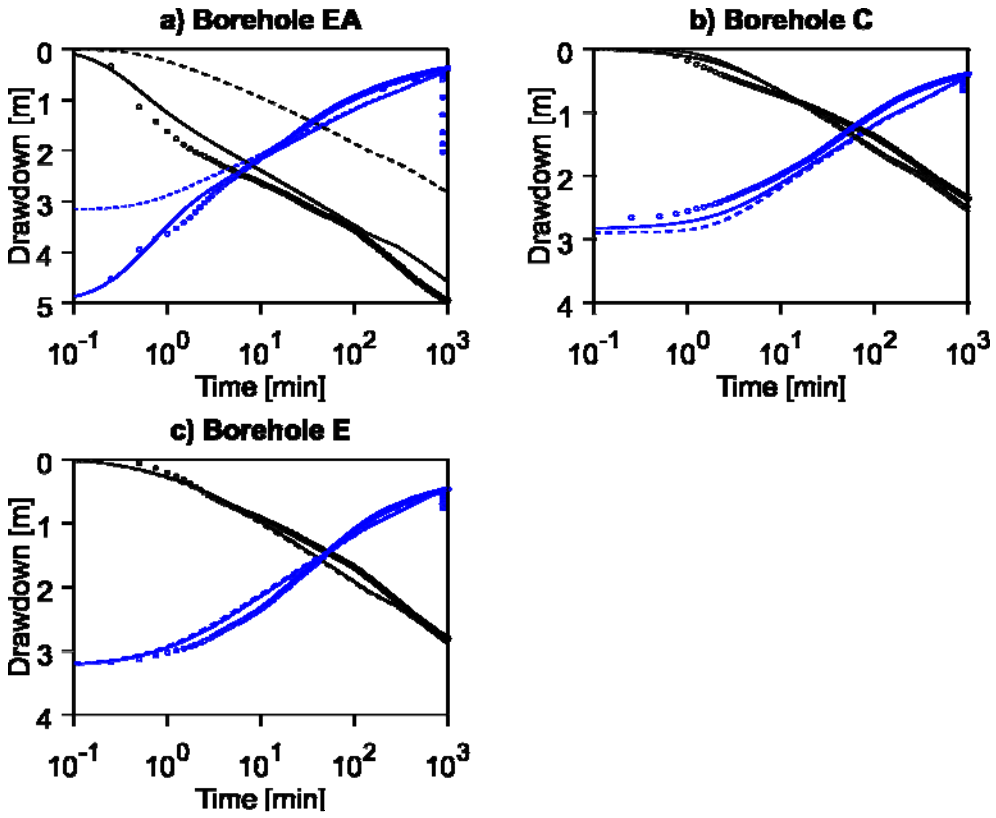



Figure 10. Comparison between observed (circles), simulated (solid lines), and predicted (dashed lines) time-drawdown curves. Black colours indicate the abstraction phase; blue colours indicate the recovery phase. The simulated values correspond to model M4. Predicted values correspond to model M4-noEA (Figure 10a), M4-noBC (Figure 10b), and M4-noADE (Figure 10c).