

1 Quantifying uncertainty in predictions of groundwater levels using 2 formal likelihood methods

3 Ben Marchant^{a*}, Jonathan Mackay^a and John Bloomfield^b

4 ¹British Geological Survey, Environmental Science Centre, Keyworth, Nottingham, NG12 5GG, UK

5 ²British Geological Survey, Maclean Building, Crowmarsh Gifford, Wallingford, Oxfordshire, OX10
6 8BB, UK

7 *Corresponding author: benmarch@bgs.ac.uk

8 Abstract

9 Informal and formal likelihood methods can be used to quantify uncertainty in modelled predictions
10 of groundwater levels (GWLs). Informal methods use a relatively subjective criterion to identify sets
11 of plausible or behavioural parameters of the GWL models. In contrast, formal methods specify a
12 statistical model for the residuals or errors of the GWL model. The formal uncertainty estimates are
13 only reliable when the assumptions of the statistical model are appropriate.

14 We apply the formal approach to historical reconstructions of GWL hydrographs from four UK
15 boreholes. We test whether a model which assumes Gaussian and independent errors is sufficient to
16 represent the residuals or whether a model which includes temporal autocorrelation and a general
17 non-Gaussian distribution is required. Groundwater level hydrographs are often observed at
18 irregular time intervals so we use geostatistical methods to quantify the temporal autocorrelation
19 rather than more standard time series methods such as autoregressive models.

20 According to the Akaike Information Criterion, the more general statistical model better represents
21 the residuals of the GWL model. However, no substantial difference between the accuracy of the
22 GWL predictions and the estimates of their uncertainty is observed when the two statistical models

23 are compared. When the general model is applied, significant temporal correlation over periods
24 ranging from 3 to 20 months is evident for the different boreholes. When the GWL model
25 parameters are sampled using a Markov Chain Monte Carlo approach the distributions based on the
26 general statistical model differ from those of the Gaussian model, particularly for the boreholes with
27 the most autocorrelation. These results suggest that the independent Gaussian model of residuals is
28 sufficient to estimate the uncertainty of a GWL prediction on a single date. However, if realistically
29 autocorrelated simulations of GWL hydrographs for multiple dates are required or if the
30 distributions of the GWL model parameters are of interest, then the more general statistical model
31 should be used.

32 **Keywords**

33 Groundwater, likelihood, mixed models, formal, MCMC

34 **1. Introduction**

35 Groundwater level (GWL) hydrographs from boreholes provide valuable information about the
36 return periods and severity of past drought and flood events. There is often a need to use
37 deterministic models to extend a hydrograph record either to reconstruct GWLs prior to the drilling
38 of the borehole (Mackay et al., 2014), to interpolate GWLs on dates when they were not observed
39 (Sun et al., 2009) or to forecast future GWLs (Daliakopoulos et al., 2004). Reconstructed hydrographs
40 might assist scientists in understanding the influence of long-term anthropogenic processes such as
41 abstraction or climate change on the variation of GWLs (Shepley and Soley, 2012) and more
42 specifically on the severity of extreme events (Kidmose et al., 2013). Interpolations of incomplete
43 hydrograph records are required to compute standardised indices of GWLs (e.g. Bloomfield and
44 Marchant, 2013) that can place current GWLs in a historical context. Forecasts of GWLs might warn
45 land managers and policy makers of potential extreme events so that remediation efforts are
46 focussed appropriately (Jackson et al., 2013).

47 In all of these contexts, it is vital that the uncertainties of the modelled GWLs are quantified so that
48 land managers and scientists interpreting predictions can determine which features reflect
49 statistically significant variation in groundwater processes rather than model errors (Jackson et al.,
50 2015). Uncertainties can arise because of errors in inputs to the groundwater models such as rainfall
51 amounts, errors in the model structure, measurement errors and errors in the estimated parameters
52 of the groundwater model. Schoups and Vrugt (2010) describe two approaches for estimating the
53 parameter uncertainty of hydrological models. Both approaches lead to an ensemble of plausible
54 parameters rather than a single optimal set. With formal likelihood methods, a statistical model for
55 the residuals is specified and used to derive a likelihood function to quantify the probability that the
56 observed data would have arisen from the hydrological model with a particular parameter set. The
57 likelihood function is then used to assess which parameter sets are plausible. The framework which
58 combines a deterministic model with a statistical model of the residuals is referred to as a mixed
59 model (MM; Dobson, 1990). The predictions from the deterministic model constitute the fixed
60 effects and the predictions of the residual model are the random effects. The random effects will
61 include contributions resulting from input errors, measurement errors and model structural errors. A
62 MM can be used to predict the entire probability density function (pdf) for the property of interest
63 on a target date when it was not observed (Lessels and Bishop, 2013). This pdf is conditioned on the
64 available observations and accounts for the correlation between the random effects on the target
65 and observation dates.

66 Beven et al. (2008) note that the formal likelihood approach relies on the assumptions of the
67 statistical model and these assumptions might be inappropriate. For example, a simple model for the
68 random effects might specify that they are independent and realized from a Gaussian distribution
69 with zero mean and constant (i.e. stationary) variance. However, autocorrelated, non-Gaussian
70 errors with non-stationary variance often occur (Kuczera, 1983). Therefore, Beven et al. (2008)
71 advocate informal likelihood methods such as generalized likelihood uncertainty estimation (GLUE;
72 Beven and Freer, 2001). In the GLUE approach the likelihood function for the errors is not linked to a

73 specific error model. Instead, metrics such as the proportion of the variance of the observations that
74 is explained by the hydrological model, are used to assess the plausibility of a particularly set of
75 parameters. Thus the GLUE approach is free from assumptions but a somewhat subjective choice of
76 likelihood function is required. The identified set of plausible parameter values can be used to
77 generate multiple modelled reconstructions of the property of interest. The between-reconstruction
78 variability corresponds to the contribution of parameter uncertainty to the total uncertainty.
79 However, there is no model to predict the other contributions such as input uncertainty and model
80 structural errors.

81 Schoups and Vrugt (2010) respond to the concerns of Beven et al. (2008) by generalizing their
82 random effects models to accommodate non-Gaussian variation, temporal correlation and non-
83 stationary variances. They assume that their random effects are realized from a skew exponential
84 power (SEP) distribution which specifies the skewness and kurtosis of the residuals independently.
85 The SEP distribution permits more general non-Gaussian variation than a transformation of the
86 observed data (e.g. Box and Cox, 1964). The variance of the residuals is permitted to vary according
87 to streamflow and the temporal correlation is represented by autoregressive time series models
88 (Chandler and Scott, 2011). Schoups and Vrugt (2010) demonstrated their approach on rainfall
89 runoff models in both humid and arid basins. They used Bayesian uncertainty methods to sample
90 plausible sets of parameters for both the fixed and random effects models. For both basins, the
91 observed flow data had a very heavy tail which resulted from the large and rapid response of the
92 flow to large storm events. Schoups and Vrugt (2010) found that their generalized non-Gaussian
93 model led to larger likelihoods than a model which assumed independent and Gaussian random
94 effects. Hence the assumption of independent and Gaussian residuals was not appropriate. The
95 generalized random effects model did not improve the model predictions. In fact, the mean squared
96 errors were smaller for the independent Gaussian model because the generalized model led to more
97 emphasis being placed on accurately estimating the low rather than large flows. The generalized
98 model did however lead to large improvements in the estimates of the uncertainty of the model

99 predictions. Also, the inappropriate Gaussian model led to quite different distributions of plausible
100 parameters than were realized from the generalized model.

101 The formal and informal likelihood approaches are also applicable to groundwater models. Jackson
102 et al. (2016) use the GLUE methodology to assess the uncertainty of model reconstructions of
103 groundwater levels at six boreholes in the UK. The reconstructions are generated using the Aquimod
104 conceptual model (Mackay et al., 2014) and the Nash Sutcliffe Efficiency (NSE) score (Nash and
105 Sutcliffe, 1970) is used to decide which sets of parameters are plausible. Von Asmuth and Bierkens
106 (2005), Mirzavand and Ghazavi (2015) and Peterson and Western (2014) all specify a formal
107 statistical model for the residuals from their models of GWLs. These statistical models are based on a
108 Gaussian distribution but they account for temporal autocorrelation amongst the residuals.

109 In this paper we quantify the uncertainty of Aquimod reconstructions for GWLs in four English
110 boreholes using a formal likelihood approach similar to Schoups and Vrugt (2010) and we discuss the
111 relative suitability of the formal and informal approaches for quantifying the uncertainty of UK
112 groundwater models. We also explore whether the assumption of independent and Gaussian
113 residuals is suitable for a formal model of GWLs in this context or whether a more general model is
114 required. Groundwater hydrographs tend to be less heavy tailed than river hydrographs since, in
115 effect, the hydrogeological system acts as a filter which temporally smooths the effects of intense
116 storm events.

117 One modification to the approach Schoups and Vrugt (2010) is necessary. The autoregressive
118 models which they use to represent temporal autocorrelation amongst the residuals are well-suited
119 to data observed at regular intervals but they cannot be applied to irregularly sampled time series.
120 Until the relatively recent installation of automated telemetry in some groundwater boreholes,
121 GWLs tended to be recorded at irregular time intervals (Environment Agency, 2014) according to
122 factors such as the availability of staff to visit the borehole and conduct a dip-test. Von Asmuth and
123 Bierkens (2005) recognised this problem and suggested a continuous approximation to the

124 autoregressive model. This approach was adopted by Peterson and Western (2014). In the more
125 general hydrological context, Chandler and Scott (2011) recommend that the temporal correlation
126 amongst irregularly sampled water levels is represented by a variogram. Variograms are more
127 commonly associated with spatial analyses (Webster and Oliver, 2007) and they describe how the
128 expected squared difference between a pair of observations varies according to the lag between the
129 observation sites or times. Chandler and Scott (2013) use the method of moments to estimate their
130 variograms but this estimator is not immediately compatible with formal likelihood functions.
131 Therefore, we consider model-based variogram estimators (Diggle and Ribeiro, 2007) that use
132 formal likelihood functions. Lessels and Bishop (2013) used a linear MM with an exponential
133 variogram to represent irregularly measured water quality parameters from two catchments in
134 southeast Australia. We represent the variograms by the flexible four parameter Matérn function
135 which generalises many commonly used variogram functions such as the exponential model
136 (Marchant and Lark, 2007).

137 **2. A mixed model for groundwater levels**

138 We represent the GWLs observed at times $t = t_1, t_2, \dots, t_n$ by a MM:

$$z(t_i) = m(t_i|\boldsymbol{\beta}_f) + r(t_i|\boldsymbol{\beta}_r), \quad (1)$$

139 where $m(t_i|\boldsymbol{\beta}_f)$ is the deterministic model prediction or fixed effect at time t_i and $r(t_i|\boldsymbol{\beta}_r)$ is the
140 random effect or residual at time t_i . The $\boldsymbol{\beta}_f$ and $\boldsymbol{\beta}_r$ are the calibrated parameters for the fixed and
141 random effects respectively. For brevity, we henceforth denote $z(t_i)$, $m(t_i|\boldsymbol{\beta}_f)$ and $r(t_i|\boldsymbol{\beta}_r)$ by
142 z_i , m_i and r_i respectively.

143 *2.1 The fixed effects model*

144 Mackay et al. (2014) reviewed the types of models used to simulate GWLs. They distinguished ‘black
145 box’ methodologies such as statistical transfer functions (e.g. Jakeman et al., 2006) from process-
146 driven models based on simplifications of physical laws of fluid dynamics (e.g. Shepley and Soley,

147 2012). They noted that the process-driven models can relate the variation of GWLs to
148 hydrogeological properties but that the calibration of these models is complex and requires more
149 data than are typically available. In contrast, black box methodologies are more easily calibrated but
150 they provide little insight into the controls on GWLs. Mackay et al. (2014) proposed Aquimod, a
151 conceptual lumped parameter model, as a compromise approach. Rather than starting from basic
152 physical principles, such conceptual models contain simpler representations of the components of
153 the hydrogeological system. We use Aquimod as the fixed effects of our MM.

154 Aquimod includes simple conceptual representations of soil drainage, the transfer of water through
155 the unsaturated zone and groundwater flow. The soil zone is represented as a bucket which receives
156 water from rainfall and releases water through evapotranspiration and drainage into the
157 unsaturated zone. Measured rainfall and potential evapotranspiration amounts are inputs to the
158 model. Model parameters control the nonlinear relationship between the rate of evapotranspiration
159 and the soil moisture and the rate at which water drains from the soil. A parametric transfer
160 function is used to represent the rate of vertical flow through the unsaturated zone into the
161 saturated zone which is represented by a layered rectangular block of aquifer through which water
162 flows horizontally. Water discharges from a layer via an outlet at its base. The discharge rate is
163 dependent on the layer permeability and the hydraulic gradient. The former is controlled using a
164 hydraulic conductivity parameter and the latter is controlled by outlet elevation and aquifer length
165 parameters. The number of layers is selected for each borehole to lead to the best possible match
166 between observed and modelled GWLs.

167 This model is easier to calibrate than a more complex process-based model and more easily
168 interpretable than a black-box model. An implementation of Aquimod with three layers in the
169 saturated zone has a total of 16 parameters but eight of these can be estimated from available
170 information about the catchment. The remaining eight parameter values constitute our fixed effects
171 parameter vector β_f . These parameters are Z_r (mm) the maximum rooting depth of vegetation, p

172 the water depletion factor of vegetation, λ the scale parameter of a Weibull function which
 173 describes the rate of recharge, k the shape parameter for the Weibull function, S (%) the aquifer
 174 storage coefficient and k_i $i = 1,2,3$ (m d^{-1}) the hydraulic conductivity for each layer of the saturated
 175 zone. Mackay et al. (2014) estimated these parameters by finding the values which led to the largest
 176 Nash-Sutcliffe Efficiency (NSE) score. The NSE score is measure of the proportion of variance which
 177 has been explained by the model and is defined as:

$$\text{NSE} = 1 - \frac{\sum_{i=1}^n \{z_i - m_i\}^2}{\sum_{i=1}^n \{z_i - \bar{z}\}^2}, \quad (2)$$

178 where \bar{z} is the mean of the observed values. Jackson et al. (2016) set a threshold of 0.5 on the NSE
 179 when deciding which parameter vectors were plausible.

180 *2.2 The random effects model*

181 We assume that the random effects, r_i , are a realization of a multivariate random function with zero
 182 mean. A simple random effects model would assume that the random function is Gaussian with
 183 fixed variance and that the realizations from this function are independent. However, this model
 184 might not be sufficiently flexible to represent observed GWLs (Bloomfield and Marchant, 2013). A
 185 more general random function can be described in terms of its marginal distributions and its
 186 dependence structure or copula (e.g. Bárdossy and Li, 2008). A marginal distribution is the pdf for a
 187 random function, in our case the random effects, at a single time. It does not take any account of the
 188 random effects at other times. In this paper, we assume that the residuals at each time are realized
 189 from the same marginal distribution with density $f(r)$ and cumulative distribution function (cdf)
 190 $F(r)$. Therefore, if an appropriate marginal distribution is specified, the set of $u_i = F(r_i)$ quantile
 191 values should be a sample from a uniform distribution bounded by zero and one. The copula
 192 describes the correlation between the u_i . If we assume a Gaussian copula and denote the cdf of a
 193 standardised Gaussian distribution by $\Phi_{0,1}$ then $\mathbf{a} = (a_1, \dots, a_n)$, where $a_i = \Phi_{0,1}^{-1}(u_i)$, is a

194 realization of a multivariate Gaussian distribution where each marginal has zero mean and unit
 195 variance and the a_i are linearly correlated with correlation matrix \mathbf{C} .

196 The log-likelihood of a multivariate random function with marginal distribution $f(r)$, Gaussian
 197 copula and correlation matrix \mathbf{C} can be written (Kazianka and Pilz, 2010; Marchant et al., 2011)

$$198 \quad l = -\frac{1}{2} \log |\mathbf{C}| + \frac{1}{2} \mathbf{a}^T (\mathbf{I}_n - \mathbf{C}^{-1}) \mathbf{a} - \sum_{i=1}^n \log[f(r_i)], \quad (3)$$

199 where \mathbf{I}_n is the identity matrix of length n .

200 We relax the standard assumption of independent Gaussian residuals by calibrating random effects
 201 models with more general marginal distributions and correlation matrices. Our choice of marginal
 202 distribution is the asymmetric exponential power (AEP) distribution (Figure 1). This has density (Zhu
 203 and Zinde-Walsh, 2009):

$$f(x) = \begin{cases} \left(\frac{\alpha}{\alpha^*}\right) \frac{1}{\sigma} K_{\text{EP}}(p_1) \exp\left(-\frac{1}{p_1} \left|\frac{x-\mu}{2\alpha^*\sigma}\right|^{p_1}\right) & \text{if } x \leq \mu \\ \left(\frac{1-\alpha}{1-\alpha^*}\right) \frac{1}{\sigma} K_{\text{EP}}(p_2) \exp\left(-\frac{1}{p_2} \left|\frac{x-\mu}{2(1-\alpha^*)\sigma}\right|^{p_2}\right) & \text{if } x > \mu. \end{cases} \quad (4)$$

204 where μ is the location parameter, $\sigma > 0$ is the scale parameter, $\alpha \in (0,1)$ is the skewness
 205 parameter, $p_1 > 0$ and $p_2 > 0$ are the left and right tail parameters, $K_{\text{EP}}(p) = 1/[2p^{1/p}\Gamma(1 +$
 206 $1/p)]$ is the normalizing constant, Γ is the Gamma function and $\alpha^* = \alpha K_{\text{EP}}(p_1)/[\alpha K_{\text{EP}}(p_1) +$
 207 $(1-\alpha)K_{\text{EP}}(p_2)]$. Figure 1 illustrates how the skewness and the decay of the left and right tails of
 208 the pdf are controlled by α, p_1 and p_2 . When $p_1 = p_2$ the AEP distribution reduces to an alternative
 209 parameterisation of the SEP distribution used by Schoups & Vrugt (2010). Zhu and Zinde-Walsh also
 210 derive expressions for the AEP cdf in terms of the Gamma cdf $G(x, \gamma)$:

$$F(x) = \begin{cases} \alpha \left[1 - G\left(\frac{1}{p_1} \left(\left|\frac{x-\mu}{2\alpha^*\sigma}\right|\right)^{p_1}, \frac{1}{p_1}\right)\right] & \text{if } x \leq \mu \\ \alpha + (1-\alpha) G\left(\frac{1}{p_2} \left(\left|\frac{x-\mu}{2(1-\alpha^*)\sigma}\right|\right)^{p_2}, \frac{1}{p_2}\right) & \text{if } x > \mu \end{cases}, \quad (5)$$

211 the quantile function of the AEP:

$$F^{-1}(v) = \begin{cases} \mu - 2\sigma\alpha^* \left[p_1 G^{-1} \left(1 - \frac{v}{\alpha}, \frac{1}{p_1} \right) \right]^{\frac{1}{p_1}} & \text{if } v \leq \mu \\ \mu + 2\sigma(1 - \alpha^*) \left[p_2 G^{-1} \left(1 - \frac{1-v}{1-\alpha}, \frac{1}{p_2} \right) \right]^{\frac{1}{p_2}} & \text{if } v > \mu \end{cases}, \quad (6)$$

212 and demonstrate that the expectation of an AEP distributed random variable X is:

$$213 \quad E(x) = \mu + \frac{\sigma}{B} \left[(1 - \alpha)^2 \frac{p_2 \Gamma(2/p_2)}{\Gamma^2(1/p_2)} - \alpha^2 \frac{p_1 \Gamma(2/p_1)}{\Gamma^2(1/p_1)} \right], \quad (7),$$

214 where $B = \alpha K_{EP}(p_1) + (1 - \alpha) K_{EP}(p_2)$.

215 Schoups and Vrugt (2010) use autoregressive models to determine the correlation between random
 216 effects. However, this approach cannot be used when the GWLs are observed at irregular time
 217 intervals. Instead, we use a model-based geostatistical approach (Diggle and Ribeiro, 2007) and
 218 calculate the entries of \mathbf{C} using an authorized parametric function which ensures that \mathbf{C} is positive
 219 definite. Many such authorized functions exist (Webster and Oliver, 2007) including the exponential
 220 model which is a continuous equivalent to an autoregressive model of order 1. We choose the more
 221 general Matérn function:

$$\mathbf{C}_{i,j} = \begin{cases} 1 & \text{if } |t_i - t_j| = 0 \\ \frac{1}{2^{\nu-1} \Gamma(\nu)} \left(\frac{|t_i - t_j|}{\rho} \right)^\nu K_\nu \left(\frac{|t_i - t_j|}{\rho} \right) & \text{if } |t_i - t_j| > 0 \end{cases} \quad (8)$$

222 to express entry i,j of the covariance as a function of $|t_i - t_j|$, the time separating the two
 223 observations. The Matérn function has two parameters, namely the distance parameter ρ and the
 224 smoothness parameter ν . The smoothness parameter controls the rate of decay of the function for
 225 small lags (Figure 2). When $\nu = 0.5$ the Matérn function is equal to the exponential function. If we
 226 select μ to ensure that $E(r) = 0$ then our general model of the residuals has six parameters to be
 227 calibrated i.e. $\boldsymbol{\beta}_r = (\sigma, \alpha, p_1, p_2, \rho, \nu)$.

228 *2.3 Calibration of the mixed model*

229 Our general MM has a total of 14 parameters which must be estimated or calibrated on observed
230 GWLs. The maximum likelihood (ML) estimator uses a numerical optimization algorithm to find the
231 values of these parameters that maximizes the log-likelihood function (Eqn. 3) for the calibration
232 data. It is also possible to compare different model structures by comparing their likelihoods. For
233 example, one might wish to consider whether an Aquimod fixed effects model with a three layer
234 saturated zone is a significantly better fit to the data than a model with only two layers.
235 Alternatively, one might consider whether a random effects model with a Matérn correlation
236 function is superior to one that uses an exponential model. This can be achieved by fitting each
237 model by ML and then comparing their Akaike Information Criterion (AIC; Akaike, 1973):

$$AIC = 2k - 2l, \quad (14)$$

238 for the maximized log-likelihood l from Equation 3. Here, k is the total number of parameters in the
239 model. If too many parameters are included in a model it might be over-fitted. This means that the
240 model matches the intricacies of the calibration data very closely but is less suitable for representing
241 independent validation data. The model with the lowest AIC is thought to be the best compromise
242 between complexity and quality of fit to the data (Webster and Oliver, 2007). Alternative
243 information criteria such as the Bayesian Information Criterion (BIC, Marshall et al., 2005) do exist.
244 However, the formula for the BIC includes the number of observations. When the observations are
245 highly correlated some adjustment of this term will be required.

246 The ML estimator is a frequentist method that looks for the single set fixed parameter values that
247 generated the observed data (Minasny et al., 2011). In reality, such a set of parameters rarely exist.
248 Deterministic models tend to approximate the complexities of environmental systems. Even if an
249 optimal deterministic model existed, it is highly unlikely that sufficient calibration data would be
250 available to uniquely identify the parameters of this model. Indeed, many deterministic models
251 include state variables that are unmeasurable. Therefore a number of models are likely to be
252 suitable to represent the environmental system (Beven and Binley, 1992).

253 In a Bayesian analysis the model parameters are treated as probabilistic variables. Our knowledge of
254 the parameter values prior to collecting any data is expressed as a prior distribution. Then the
255 observations are used to update these priors and to form a posterior distribution of the GWLs which
256 combines our prior knowledge with the information that could be inferred from the observations.
257 The posterior distribution can be sampled using a Markov Chain Monte Carlo (MCMC) approach
258 (Diggle and Ribeiro, 2007). This is an iterative method which moves between behavioural or
259 plausible parameter vectors according to the corresponding values of the log-likelihood function.
260 The parameter vector is randomly perturbed and the Metropolis-Hastings algorithm (Hastings, 1970)
261 is used to decide if the parameter set is behavioural. The MCMC approach is computationally
262 demanding and the perturbations of the parameter vector must be carefully controlled to ensure
263 that the sample is representative of the behavioural parameter set. Until recently, these challenges
264 might have prevented the use of the approach to estimate all of the parameters of a deterministic
265 GWL model. However, Vrugt et al. (2008) have developed the DiffeRential Evolution Adaptive
266 Metropolis (DREAM) algorithm which permits efficient MCMC sampling in high dimensional
267 parameter spaces and automatically selects effective perturbations of the parameter vector.
268 Minasny et al. (2011) demonstrated how this algorithm could be applied in conjunction with
269 geostatistical models and Minasny et al. (2013) described how it could be extended to MMs that
270 included nonlinear fixed effects models. The MCMC sample can also be used to assess whether the
271 parameters of the fixed effects model are identifiable. This concept is formally defined by Renard et
272 al. (2010). A parameter is non-identifiable if the observed data do not provide any information about
273 that parameter. In this case the posterior distribution of the parameter is no more certain than the
274 prior distribution.

275 *2.4 Prediction using the mixed model*

276 The calibrated mixed model can be used to make a prediction of the pdf of the residuals or to
277 generate simulations of the residuals at a set of times where the GWL was not observed. These

278 predictions and simulations are conditional on the observations that are available. In our discussion
 279 above we demonstrated that $\mathbf{a} = (a_1, \dots, a_n)$, where $a_i = \Phi_{0,1}^{-1}(u_i)$ is a realization of a multivariate
 280 Gaussian random function. The kriging predictor (Webster & Oliver, 2007) can be used to predict a
 281 at a set of q target times $\mathbf{T} = (T_1, \dots, T_q)$ when it was not observed. We denote these predictions by
 282 \mathbf{a}_T . The length q vector of expectations \mathbf{e}_T and the $q \times q$ covariance matrix \mathbf{v}_T of \mathbf{a}_T are:

$$\mathbf{e}_T = \mathbf{C}_{T0}\mathbf{C}^{-1}\mathbf{a}, \quad (9)$$

$$\mathbf{v}_T = \mathbf{C}_{TT} - \mathbf{C}_{T0}\mathbf{C}^{-1}\mathbf{C}_{T0}^T, \quad (10)$$

283 where \mathbf{C}_{T0} is the $q \times n$ matrix of correlations between the residuals on the target times and the
 284 observed residuals and \mathbf{C}_{TT} is the correlation matrix for the residuals at the target times. The LU
 285 method (Webster & Oliver, 2007) can be used to generate simulations of \mathbf{a}_T . These can be
 286 transformed to simulations of the residuals:

$$\mathbf{r}_T = F^{-1}[\Phi_{0,1}(\mathbf{a}_T)]. \quad (11)$$

287 The correlation between the elements of each of the realizations will be consistent with the
 288 calibrated model. Alternatively it is possible to predict the pdf of the residual for a single target time
 289 conditional on the observed residuals \mathbf{r} by calculating:

$$f(r^*|\mathbf{r}, \boldsymbol{\beta}_r) = \frac{f(r^*) \times \phi_{e_T, \sqrt{v_T}}(a^*)}{\Phi_{0,1}(a^*)}, \quad (12)$$

290 for the range of plausible values of r^* . Here, $a^* = \Phi_{0,1}^{-1}[F(r^*)]$, $\phi_{e,b}$ is the pdf of a Gaussian
 291 distribution with mean e and standard deviation b and e_T and v_T are the expectation and variance
 292 of the kriged prediction on this single date. Note that if the residuals on the target date are
 293 independent of the conditioning observed residuals or if no conditioning observations are included
 294 in the prediction, then $e_T = 0$, $v_T = 1$ and Eqn. 12 reduces to $f(r^*|\mathbf{r}, \boldsymbol{\beta}_r) = f(r^*)$.

295 If the MM has been calibrated by the MCMC approach then an ensemble of independent
296 realisations of the parameter vector will have been sampled. The pdf of the residual conditional on
297 the observed GWLs is then equal to Eqn. 12 averaged across the parameter vectors. This pdf
298 accounts for the uncertainty in estimating the fixed and random effects parameters and the residual
299 errors of the fixed effects model.

300 *2.5 Validation of the mixed model*

301 The NSE score (Eqn. 2) is commonly used as a criterion to validate hydrological models. However,
302 Thyer et al. (2009) note that the NSE score is only a measure of the accuracy of the predictions and it
303 cannot be used to confirm that the assumed distribution of the random effects is consistent with the
304 observed data. Therefore, Thyer et al. (2009) recommend the use of the predictive QQ plot. If the
305 calibrated MM is used to predict the i th observed GWL then the p-value of the observed value is
306 equal to:

$$p_i = \Phi_{0,1} \left(\frac{\tilde{a}_i - e_i}{\sqrt{v_i}} \right), \quad (13)$$

307 where \tilde{a}_i is the observed value of a at time i (i.e. $\tilde{a}_i = \Phi_{0,1}^{-1}[F(\tilde{r}_i)]$), and e_i , v_i are the expectation
308 (Eqn. 9) and variance (Eqn. 10) of the kriged prediction of a_i . If the observed GWL is a realization of
309 the MM then p_i is a realization of a uniform distribution on $[0,1]$. A QQ plot is constructed by
310 calculating p_i for a large number of observations. The p_i are sorted and plotted against the
311 theoretical p-values or the cdf of the uniform distribution (i.e. evenly spaced values between zero
312 and one). If all of the points of the QQ plot lie on the 1:1 line, the MM predictive distribution agrees
313 exactly with the observations. If all the points lie above (or alternatively, below) the 1:1 line then the
314 GWLs are under (over) predicted. If the points lie below (above) the 1:1 line for small theoretical p-
315 values and above (below) the 1:1 line for large theoretical p-values then the predictive uncertainty
316 of the MM is under (over) estimated.

317 **3. Methods**

318 All of the computations were conducted using Matlab (Mathworks, 2014) and the Matlab
319 implementation of the DREAM algorithm (Vrugt, 2016) was used to perform the MCMC analyses.

320 *3.1 Borehole and Meteorological Data*

321 We estimated MMs for the monthly records from four boreholes considered by MacKay et al.
322 (2014). These boreholes are named Chilgrove House (540 observations), Hucklow South (440
323 observations), Lower Barn Cottage (368 observations) and Skirwith (326 observations) and they are
324 set in chalk, limestone, lower greensand and sandstone respectively. Three of these boreholes were
325 considered by Jackson et al. (2016) when they used GLUE to quantify uncertainty in a GWL model.
326 MacKay et al. (2014) give full details about the characteristics and setting of the boreholes. The GWL
327 records were extracted from the UK National Groundwater Archive (National Groundwater Level
328 Archive, 2013). Monthly GWLs from the boreholes are shown in Figure 3. Strong seasonal patterns
329 are evident in all of the hydrographs and the Lower Barn Cottage and Skirwith hydrographs are
330 considerably smoother than the other two. We follow MacKay et al. (2014) and use the first half of
331 these time series for calibration of our MMs and the second half for validation. There are 18 missing
332 observations from Skirwith during the validation period.

333 The monthly precipitation data required as an input to AquMod were extracted from the Centre for
334 Ecology and Hydrology's CERF 1km gridded precipitation dataset which is derived from UK
335 Meteorological Office data (Keller et al., 2005). The monthly potential evapotranspiration time series
336 were extracted from the Meteorological Office Rainfall and Evaporation Calculation System (Field,
337 1983). These are based on a modified version of the Penman-Monteith equation (Monteith and
338 Unsworth, 2008).

339 *3.2 Model calibration and analyses*

340 We calibrated a series of MMs with increasingly complex random effects using the ML estimator.
341 The initial models assumed that the random effects were independent and realized from a Gaussian
342 distribution with constant variance. This Gaussian model was then generalized to include temporal
343 correlation described firstly by exponential and then by Matérn covariance functions. The
344 independent, exponential and Matérn covariance models were then used in conjunction with the
345 AEP distribution. The structure of the fixed effects models were identical to the Aquimod models
346 used by MacKay et al. (2014). The Chilgrove House and Hucklow South models had three saturated
347 zone structures whereas there were only two saturated zone structures for Lower Barn Cottage and
348 Skirwith. The AIC was calculated for each calibrated MM. For each borehole, the MM with the lowest
349 AIC was used to predict the GWLs during the validation period and to calculate the uncertainty of
350 these predictions. Predictive QQ plots were calculated for the calibration observations (without
351 conditioning data) and the validation observations.

352 For each borehole, the MMs were recalibrated using the DREAM MCMC approach. All of the
353 Aquimod parameters were assumed to have uniform prior distributions. The bounds on these
354 parameters were identical to the parameter ranges considered by MacKay et al. (2014). The MCMC
355 was iterated 600 000 times. Initial runs of the algorithm indicated that around 15 000 iterations
356 were required before the Gelman-Rubin convergence diagnostic (Vrugt et al., 2009) was consistently
357 less than 1.2 indicating that the MCMC had converged to the plausible portion of the parameter
358 space. There was evidence of some correlation between sampled parameter vectors separated by
359 up to 100 iterations. We therefore conservatively discarded the first 100 000 parameter vectors and
360 only selected every 500th on the remaining vectors to yield an ensemble of 1 000 parameter vectors
361 which we treated as if they were independent samples of the parameter set. The validation
362 procedures that had been applied to the ML estimates were then repeated for the ensembles of
363 MCMC parameter estimates. We also used the MCMC ensembles to assess which of the Aquimod
364 parameters were identifiable.

365 4. Results

366 4.1 Maximum likelihood estimation of the mixed models

367 Table 1 shows the AIC values for the ML estimates of the different MMs for each borehole. In each
368 case, the inclusion of the AEP rather than Gaussian random effects and the inclusion of
369 autocorrelated random effects led to a decrease in the AIC. In contrast the NSE scores (Table 2) were
370 largely unchanged as MM complexity was increased. Indeed, in the case of Lower Barn Cottage there
371 is a sharp decrease in NSE when the AEP random effects with exponential correlation function are
372 generalised to a Matérn correlation function. For three of the four boreholes the lowest AIC was
373 achieved with an exponential covariance function but at Lower Barn Cottage there was sufficient
374 improvement in likelihood to justify the use of a Matérn function.

375 The predicted pdfs of the random effects based on these best fitting MMs varied in terms of the
376 magnitude and direction of their skew and the rate of decay of each tail (Figure 4). However, the
377 observed residuals from the fixed effects model were generally consistent with these predicted pdfs.
378 The best fitting models also differ in terms of their autocorrelation functions (Figure 5). At Chilgrove
379 House and Hucklow South temporal autocorrelation is only evident for time lags of less than 5
380 months whereas for Skirwith there is correlation for lags up to 20 months and for Lower Barn
381 Cottage, observations separated by well over 20 months are autocorrelated.

382 The predictions of GWLs for the four sites during the validation period (Figure 6) followed the same
383 pattern of peaks and troughs as the observed values and the observations were generally within the
384 90% confidence limits of the predictions. The predictive QQ plots for the calibration data at
385 Chilgrove House, Hucklow South and Skirwith were all reasonably close to the 1:1 line for both the
386 Gaussian independent random effects (Figure 7) and the best fitting general random effects model
387 (Figure 8). However, the corresponding plots for Lower Barn Cottage were further from the 1:1 line,
388 particularly for the more general model where the curve was consistently above the 1:1. This

389 indicates that there is systematic under prediction of GWLs and is consistent with the relatively poor
390 NSE score for this site. We suspected that there were too few observations to accurately estimate all
391 of the components of the general MM for Lower Barn Cottage. Also, the observations that were
392 available were highly correlated. Therefore, we re-calibrated the MM for this site using both the
393 original calibration and validation observations and saw a marked improvement in the QQ plot
394 (Figure 8e). At Chilgrove House, the QQ plot for the validation data was also close to the 1:1 line for
395 both the Gaussian and general random effects models. There was some moderate under estimation
396 of the uncertainty of the validation predictions using the general random effects model at Hucklow
397 South and slightly more severe over-estimation of the uncertainty at Skirwith. The validation QQ
398 plots for both the Gaussian and best fitting random effects models were both relatively poor for
399 Lower Barn Cottage.

400 *4.2 MCMC estimation of parameters*

401 The ensembles drawn from the MCMC samplers indicate that the Aquimod parameters (Figure 9)
402 and parameters of the AEP distribution (Figure 10) are generally identifiable for each borehole. The
403 parameters are confined to a range that is less than that of the prior distribution. Note that the
404 range on the x-axis in these plots is identical to the range of the prior uniform distribution of the
405 parameter. However, the spread of the posterior realizations of k , the shape parameter of the
406 Weibull distribution within Aquimod, is almost as wide as the prior distribution for all of the
407 boreholes. Closer inspection of the MCMC ensembles revealed that this parameter is highly
408 correlated to λ , the scale parameter of the Weibull distribution, and this relationship explains the
409 identifiability issue. At Lower Barn Cottage and Skirwith the identifiability of many of the Aquimod
410 parameters improves when the entire observation record, rather than half of it, is used for
411 calibration.

412 The posterior ensembles from the Gaussian independent random effects model (red histograms in
413 Figure 8) and the best fitting random effects model (grey histograms) are relatively similar for

414 Chilgrove House and Hucklow South. However, for Lower Barn Cottage and Skirwith, there are
415 marked differences between the posterior distributions of the parameters.

416 The MCMC ensembles of random effects parameters were used to estimate the uncertainty of the
417 correlation functions (Figure 5). These plots illustrate that there is significant temporal auto-
418 correlation amongst the random effects at the $p=0.05$ level for more than 2 months at Chilgrove
419 House, more than 5 months at Hucklow South and more than 20 months at Lower Barn Cottage and
420 Skirwith. Figure 11 shows histograms of NSE values achieved by each parameter vector within the
421 MCMC ensembles for each borehole. For Chilgrove House and Hucklow South these values are fairly
422 tightly clustered around the maximum. For Lower Barn Cottage and Skirwith the NSE values are
423 more variable indicating that the proportion of GWL variation explained by the fixed effects models
424 varies between different parameter vectors within the ensemble. The NSE values for these two
425 boreholes do become more clustered close to the maximum when the entire observation record is
426 used to calibrate the model.

427 **5. Discussion**

428 *5.1 Overview*

429 We have demonstrated how MMs can be used to represent the temporal variation of GWLs at
430 specific boreholes and to predict these GWLs on dates where they were not measured. These
431 predictions can be used to reconstruct hydrographs for times prior to the drilling of the borehole, to
432 fill in gaps in the hydrograph through interpolation and to simulate potential future characteristics of
433 hydrographs under different climate scenarios. The MM framework is flexible in terms of the
434 deterministic model that may be included in the fixed effects and the structure of the random
435 effects. The MMs were tested using the same monthly GWL observations that had previously been
436 modelled by Mackay et al. (2014) using informal methods. However, the correlation functions
437 included in the random effects are fully compatible with the irregularly sampled hydrographs that

438 are available for many sites in the UK (Environment Agency, 2014). The uncertainty of MM
439 parameters can be accounted for by sampling them using a MCMC approach. This reveals that they
440 are generally identifiable although some parameters cannot be uniquely defined if they are strongly
441 related to each other (Renard et al., 2009).

442 *5.2 Structure of the random effects*

443 For all four boreholes the best fitting MM according to the AIC included temporal autocorrelation
444 amongst the random effects which were realized from an AEP rather than a Gaussian distribution.
445 This indicates that the residuals are inconsistent with the assumptions that they were independent
446 and realized from a Gaussian distribution. In this respect our results agree with the findings of
447 Schoups and Vrugt (2010) for rainfall runoff models. Also in common with Schoups and Vrugt (2010),
448 we found that the accuracy of the GWL predictions was not substantially improved by including the
449 more general random effects model. The NSE scores achieved in this study were a very slight
450 improvement on those recorded by Mackay et al. (2014) for the same boreholes but we suspect that
451 these improvements were wholly due to differences in the numerical methods used to minimize the
452 objective function when estimating the parameters rather than a difference in the modelling
453 approach.

454 Schoups and Vrugt (2010) found that the inclusion of their general random effects model did lead to
455 substantial improvements in the estimates of the predictive uncertainty. This did not occur for the
456 predictions from Aquimod. This difference could have arisen because the deviations of GWLs from
457 the Gaussian distribution are far less severe than for stream flows where very heavy tails result from
458 sharp responses to storm events. Indeed, the models of Schoups and Vrugt (2010) included a
459 relationship between the error variance and the flow rate but when we experimented with such
460 relationships for Aquimod (results not shown) the likelihood did not improve.

461 The QQ plots indicated that the predictive uncertainty of *AquiMod* was relatively poorly estimated
462 by the autocorrelated AEP models for the two sites with substantial temporal correlation. We
463 suspect this was because there were too few observations from these sites and those that were
464 available were too strongly correlated to accurately estimate all of the parameters of the MM. The
465 likelihood function for an auto-correlated variable is known to be particularly sensitive to the
466 correlation between close pairs of observations (Stein, 1999). It appears that when insufficient data
467 are available, the general random effects models lead to an emphasis being placed on accurately
468 estimating the autocorrelation function at the expense of the marginal distributions and fixed effects
469 parameters. Therefore poor NSE scores and QQ plots can result. The QQ plots for the boreholes with
470 strong autocorrelation were improved when the number of calibration data was doubled. The
471 number of observations required to accurately estimate all parameters of the MM will depend on
472 the complexity of the model and the amount of autocorrelation amongst the residuals. Therefore, it
473 is not possible to give general guidelines about the data requirements and fitted MMs should be
474 carefully validated to confirm their adequacy.

475 The QQ plots using independent validation data tended to be further from the 1:1 line than those
476 based on the calibration data. This could be due in part to changes in the accuracy of the rainfall
477 data over time. Jackson et al. (2016) discuss how the density of UK rain gauges varies over time.

478 There are two more substantive implications of assuming independent and Gaussian random effects
479 when estimating random effects. First, the method will fail to identify temporal autocorrelation
480 amongst the random effects. Significant temporal correlation was identified for all four boreholes
481 and for two of the boreholes this continued for ranges up to 20 months. If this temporal correlation
482 is not modelled then it will not be accounted for when using observations to condition predictions of
483 GWLs (Eqn. 12). For example, if predictions of the GWL were required one month prior to the
484 observational record, one would expect them to be correlated to the first few observations from the
485 record and predictions which ignored the autocorrelation would be suboptimal. It is also important

486 to account for temporal autocorrelation when simulating GWLs on multiple dates. If realisations of
487 the hydrograph are produced where the monthly GWLs are erroneously independent, the
488 uncertainty of the duration of events such as droughts that span multiple months will be poorly
489 estimated. The second implication of inappropriate assumptions in the random effects is that the
490 parameters of the fixed effects will be poorly estimated.

491 *5.3 Formal and informal approaches to quantifying uncertainty of groundwater levels*

492 The formal likelihood methods applied here considered the effects of parameter uncertainty and the
493 combined effects of model specification, input and measurement errors that are included in the
494 random effects. If the random effects model assumptions are appropriate then these are calculated
495 using objective statistical methods. In contrast the informal uncertainty methods as implemented by
496 Jackson et al. (2016) and Mackay et al. (2014) associate all of the predictive uncertainty with the
497 parameter uncertainty. A subjective threshold is placed on the NSE or a similar criterion to decide
498 whether a proposed model is behavioural. The plots of NSEs realized from the MCMC analysis of our
499 models (Figure 11) suggest that a single NSE threshold is unlikely to be suitable for all boreholes.
500 Indeed, the ensembles of behavioural parameters identified by Mackay et al. (2014) suggest they are
501 considerably less identifiable than those in Figure 9. However, we note that despite these
502 misgivings, the containment ratios recorded by Jackson et al. (2016) are comparable to those that
503 can be inferred from our QQ plots, that the informal methodologies can be implemented more
504 quickly than our formal likelihood approaches and that no assumptions about the structure of the
505 model errors are required.

506 *5.4 Further generalisations of the random effects models*

507 Although the random effects models applied in this paper are substantially more flexible than
508 standard independent Gaussian models there are many ways in which they could be further
509 generalised. For example, Schoups & Vrugt (2010) permitted the variability of runoff models to

510 linearly increase according to the flow. The variance of the random effects might also be permitted
511 to vary according to GWLs, seasonally or according to any relevant covariate (Marchant et al., 2009).
512 Such changes can easily be incorporated into the MM (Eqn. 1). It is possible to incorporate any
513 marginal distribution for the random effects into the copula-based framework (Eqn. 3). The fit of the
514 MM might also be improved by incorporating a non-Gaussian dependence structure into this
515 framework. Eltahir and Yeh (1999) noted that groundwater drought episodes tend to last longer
516 than flood episodes. This suggests that the correlation between successive random effects during
517 droughts might be stronger than that during floods. Such behaviour could be accommodated by
518 using a non-Gaussian and non-symmetric copula model (Bárdossy and Li, 2008) for the dependence
519 structure. Before applying any of these generalisations it will be necessary to confirm that they lead
520 to a substantial improvement in the likelihood and AIC and to confirm by validation that the
521 resultant predictive distributions are appropriate.

522 **6. Conclusions**

523 Mixed models estimated by formal likelihood methods can be used to predict GWLs and to estimate
524 the uncertainty of these predictions. In contrast to informal methods, the criterion used to estimate
525 the models is objective and based on the likelihood that the observed data would have been realized
526 from the specified model. However, these likelihoods are only appropriate if the assumptions on
527 which they are based are appropriate. Therefore it is necessary to thoroughly validate the estimated
528 MM through methods such as predictive QQ plots which assess the accuracy of the entire predictive
529 distribution rather than just the accuracy of the estimated expected GWL at each time. If the
530 validation results are poor then generalisations of the random effects model should be considered.
531 GWLs recorded a month apart can be highly correlated and therefore a substantial number of
532 observations may be required to accurately estimate all of the components of the MM. Our tests of
533 the MM on four UK GWL hydrographs indicated that the assumptions of independent and Gaussian
534 errors are unlikely to be completely appropriate. However, the application of these inappropriate

535 models did not lead to a substantial deterioration of the GWL predictions or the estimates of their
536 uncertainty. The appropriateness of the random effects model is more important in circumstances
537 where the temporal correlation of the random effects or the posterior distributions of the fixed
538 effects parameters are of interest.

539 **Acknowledgements**

540 The work described has been funded by the British Geological Survey (Natural Environment Research
541 Council) and the Natural Environment Research Council funded project 'Analysis of historic drought
542 and water scarcity in the UK: a systems-based study of drivers, impacts and their interactions'
543 (NE/L010151//). This paper is published with the permission of the Executive Director of the British
544 Geological Survey (Natural Environment Research Council).

545 **References**

546 Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: B.N.
547 Petov and F. Csaki (eds) Second International Symposium on Information Theory, Akademia Kiado,
548 Budapest, 267—281.

549 Bárdossy, A., Li, J., 2008. Geostatistical interpolation using copulas. *Water Resources Research*, 44,
550 W07412.

551 Beven, K., Binley, A., 1992. The future of distributed models – model calibration and uncertainty
552 prediction. *Hydrological Processes*, 6, 279—298.

553 Beven, K., Freer, J., 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic
554 modelling of complex environmental systems using the GLUE methodology. *Journal of Hydrology*,
555 249, 1—29.

556 Beven, K., Smith, P.J., Freer, J.E., 2008. So just why would a modeler choose to be incoherent.
557 *Journal of Hydrology*, 354, 15—32.

558 Bloomfield J.P., Marchant B.P., 2013. Analysis of groundwater drought using a variant of the
559 Standardised Precipitation Index. *Hydrology and Earth Systems Science*, 17, 4769—4787.

560 Box, G.E.P., Cox, D.R., 1964. An analysis of transformations. *Journal of the Royal Statistical Society.*
561 *Series B (Methodological)*, 26, 211—252.

562 Chandler, R., Scott, M., 2011. *Statistical Methods for Trend Detection and Analysis in the*
563 *Environmental Sciences. Statistics in Practice*, Wiley, UK.

564 Daliakopoulos, I.N., Coulibaly, P., Tsanis, I.K., 2004. Groundwater level forecasting using artificial
565 neural networks. *Journal of Hydrology*, 309, 229—240.

566 Diggle, P.J., Ribeiro, P.J., 2007. *Model-based Geostatistics*. Springer, New York.

567 Dobson, A.J., 1990. *An Introduction to Generalized Linear Models*. 2nd Edition. Chapman and Hall,
568 UK.

569 Eltahir, E.A.B., Yeh, P. J-F., 1999. On the asymmetric response of aquifer water level to floods and
570 droughts in Illinois, *Water Resources Research*, 35, 1199-1217.

571 Environment Agency. 2014. National groundwater level database for England, Environment Agency.
572 <http://data.gov.uk/data> last retrieved 26 August 2014.

573 Field, M., 1983. The meteorological office rainfall and evaporation calculation system – MORECS.
574 *Agricultural Water Management*, 6, 297—306.

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

577 Jackson, C.R., Wang, L., Pachocka, M., Mackay, J.D., Bloomfield J.P., 2016. Reconstruction of multi-
578 decadal groundwater level time series using a lumped conceptual model. *Hydrological Processes*.
579 DOI:10.1002/hyp.10850.

580 Jackson, C.R., Pachocka, M., Mackay, J., 2013. Hydrological Outlook: Outlook based on modelled
581 groundwater levels and seasonal climate forecast. British Geological Survey Open Report OR/13/046,
582 Nottingham, UK
583 (http://www.hydoutuk.net/files/2013/9566/7747/HO_methodology_Groundwater_Levels.pdf, last
584 accessed March 2016).

585 Jackson, C.R., Bloomfield, J.P., Mackay, J.D., 2015. Evidence for changes in historic and future
586 groundwater levels in the UK. *Progress in Physical Geography*, 39, 49—67.

587 Jakeman, A.J., Letcher, R.A., Norton, J.P., 2006. Ten iterative steps in development and evaluation of
588 environmental models. *Environmental Modelling and Software*, 21, 602-614.

589 Kazianka, H., Pilz, J., 2010. Copula-based geostatistical modelling of continuous and discrete data
590 including covariates. *Stochastic Environmental Research and Risk Assessment*, 661-673.

591 Keller, V., Young, A.R., Morris, D., Davies, H., 2005. Continuous Estimation of River Flows (CERF),
592 Technical Report: Estimation of Precipitation Inputs, Environment Agency R&D Project Report WD-
593 101, Centre for Ecology and Hydrology, Wallingford.

594 Kidmose, J., Refsgaard, J.C., Troldborg, L., Seaby, L.P., Escrivà, M., 2013. Climate change impact on
595 groundwater levels: ensemble modelling of extreme values. *Hydrology and Earth Systems Sciences*,
596 17, 1619—1634.

597 Kuczera, G., 1983. Improved parameter inference in catchment models: 1.Evaluating parameter
598 uncertainty, *Water Resources Research*, 19, 1151—1162.

599 Lessels, J.S., Bishop, T.F.A., 2013. Estimating water quality using linear mixed models with stream
600 discharge and turbidity. *Journal of Hydrology*, 498, 13—22.

601 Mackay, J.D., Jackson, C.R., Wang, L., 2014. A lumped conceptual model to simulate groundwater
602 level time-series. *Environmental Modelling and Software*, 61, 229—245.

603 Marchant, B.P., Lark, R.M. 2007. The Matérn variogram model: Implications for uncertainty
604 propagation and sampling in geostatistical surveys. *Geoderma*, 140, 337–345.

605 Marchant, B.P., S. Newman, S., Corstanje, R., Reddy, K.R., Osborne, T.Z., Lark, R.M., 2009. Spatial
606 monitoring of a non-stationary soil property: Phosphorus in a Florida water conservation area.
607 *European Journal of Soil Science*, 60(5), 757-769.

608 Marchant, B.P., Saby, N.P.A., Jolivet, C.C., Arrouays, D., Lark, R.M., 2011. Spatial prediction of soil
609 properties with copulas. *Geoderma*, 162, 327-334.

610 Marshall, L., Nott, D., Sharma, A., 2005. Hydrological model selection: A Bayesian alternative. *Water*
611 *Resources Research*, 45, W04418.

612 Mathworks, 2014. MATLAB and Statistics Toolbox Release 2014b, The MathWorks, Inc., Natick,
613 Massachusetts, United States.

614 Minasny, B., Vrugt, J.A., McBratney, A.B., 2011. Confronting uncertainty in model-based geostatistics
615 using Markov Chain Monte Carlo simulation. *Geoderma*, 163, 150–162.

616 Minasny, B., Whelan, B.M., Triantafilis, J., McBratney, A.B., 2013. Pedometrics research in the
617 vadose zone-Review and perspectives. *Vadose Zone Journal*, 12, 4.

618 Mirzavand, M., Ghazavi, R., 2014. A stochastic modelling technique for groundwater level
619 forecasting in an arid environment using time series methods. *Water Resources Management*, 29,
620 1315–1328.

621 Monteith, J.L., Unsworth, M.H., 2008. *Principles of Environmental Physics: Third Edition*. Elsevier,
622 London, UK.

623 Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I – A discussion
624 of principles. *Journal of Hydrology*, 10, 282–290.

625 National Groundwater Level Archive. <http://www.ceh.ac.uk/data/nrfa/data/ngla.html>, last access:
626 March 2015.

627 Peterson, T.J., Western, A.W., 2014. Nonlinear time-series modelling of unconfined groundwater
628 head. *Water Resources Research*, 50, 8330—8355.

629 Renard, B., Kavetski, D., Kuczera, G., Thyer, M., Franks, S., 2010. Understanding predictive
630 uncertainty in hydrologic modelling: The challenges of identifying input and structural errors. *Water*
631 *Resources Research*, 46, W05521.

632 Schoups, G., Vrugt, J.A., 2010. A formal likelihood function for parameter and predictive inference of
633 hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. *Water Resources*
634 *Research*, 46, W10531.

635 Shepley, M.G., Soley, R.W.N., 2012. The use of groundwater levels and numerical models for the
636 management of a layered, moderate-diffusivity aquifer. *Geological Society, London, Special*
637 *Publications*, 364, 303—318.

638 Sun, Y., Kang, S., Li, F., Zhang, L., 2009. Comparison of interpolation methods for depth to
639 groundwater and its temporal and spatial variations in the Minqin oasis of northwest China.
640 *Environmental Modelling and Software*, 24, 1163—1170.

641 Stein, M.L., 1999. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer-Verlag, New York

642 Thyer, M., Renard, B., Kavetski, D., Kuczera, G., Franks, S., Srikanthan, S., 2009. Critical evaluation of
643 parameter consistency and predictive modelling in hydrological modelling: A case study using
644 Bayesian total error analysis. *Water Resources Research*, 45, W00B14.

645 von Asmuth, J.R., Bierkens, M.F., 2005. Modelling irregularly spaced residual series as a continuous
646 stochastic process. *Water Resources Research*, 41, W12404.

647 Vrugt, J.A., ter Braak, C.J.F., Clark, M.P., Hyman, J.M., Robinson, B.A., 2008. Treatment of input
648 uncertainty in hydrologic modeling: doing hydrology backward with Markov chain Monte Carlo
649 simulation. *Water Resources Research* 44, W00B09.

650 Vrugt, J.A., ter Braak, C.J.F, Diks, C.G.H, Robinson, B.A., Hyman, J.M., Higdon, D., 2009. Accelerating
651 Markov Chain Monte Carlo simulation by differential evolution with self-adaptive randomized
652 subspace sampling. *International Journal of Nonlinear Science and Numerical Simulation*, 10, 273-
653 290.

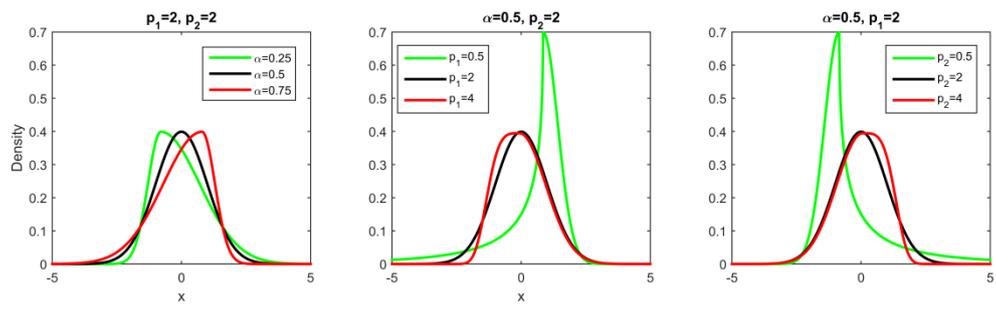
654 Vrugt, J.A., 2016. Markov Chain Monte Carlo simulation using DREAM software package: Theory,
655 concepts, and Matlab implementation. *Environmental Modelling and Software*, 75, 273-316.

656 Webster, R., Oliver, M.A., 2007. *Geostatistics for Environmental Scientists*, 2nd ed., John Wiley and
657 Sons, Chichester, UK.

658 Zhu, D., Zinde-Walsh, V., 2009. Properties and estimation of asymmetric exponential power
659 distribution. *Journal of Econometrics*, 148, 86-99.

660

661 **Figures**



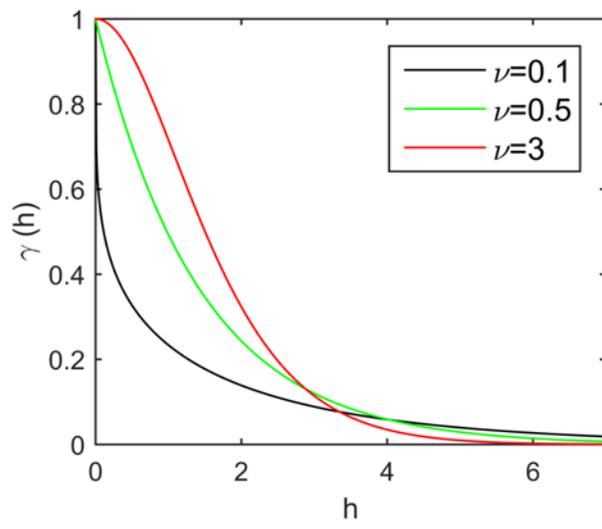
662

663 **Figure 1:** Examples of the AEP pdf for $\mu = 0, \sigma = 1$ and stated values of α, p_1 and p_2 .

664

665

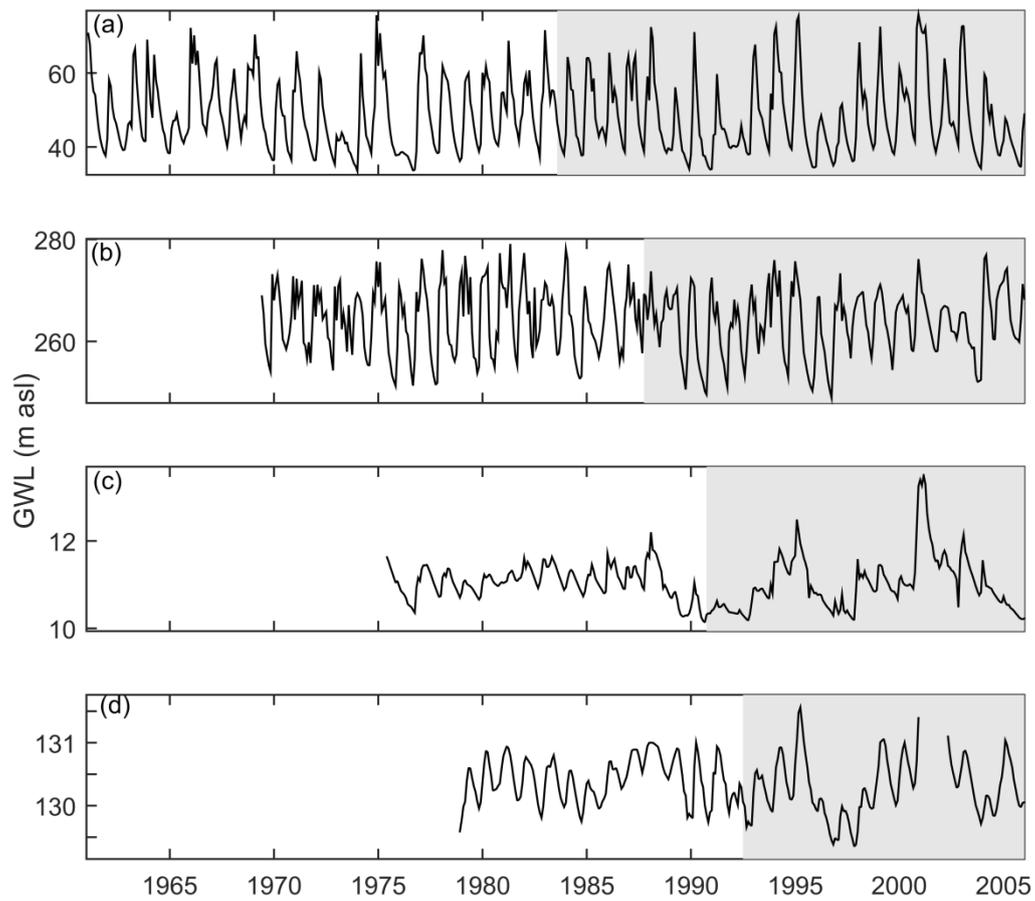
666



667

668 **Figure 2:** Examples of the Matérn covariance function with $c_1 = 1$, $\rho = 2$ and stated values of ν .

669



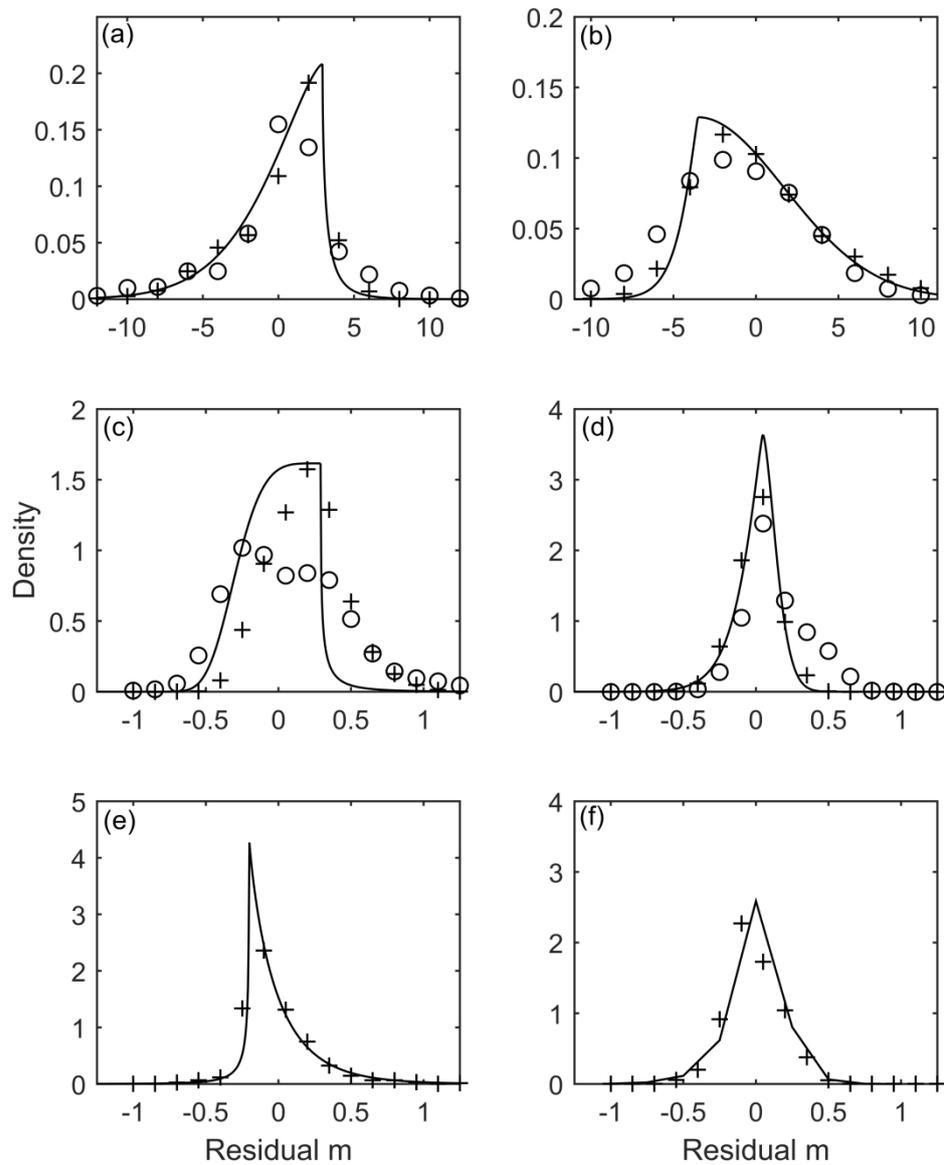
670

671

672 **Figure 3:** Observed monthly GWLs from (a) Chilgrove House, (b) Hucklow South, (c) Lower Barn

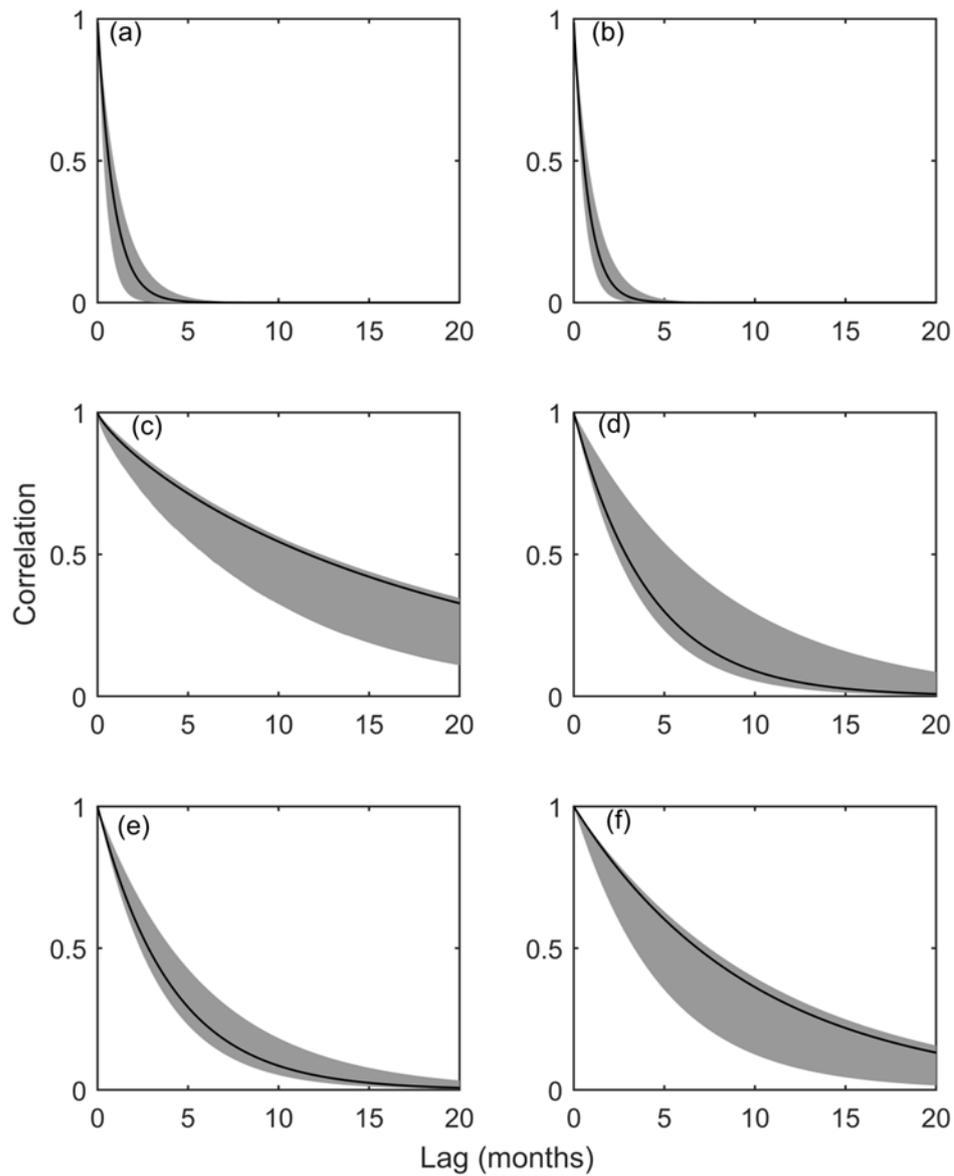
673 Cottage, (d) Skirwith. Validation period is shaded grey.

674



675

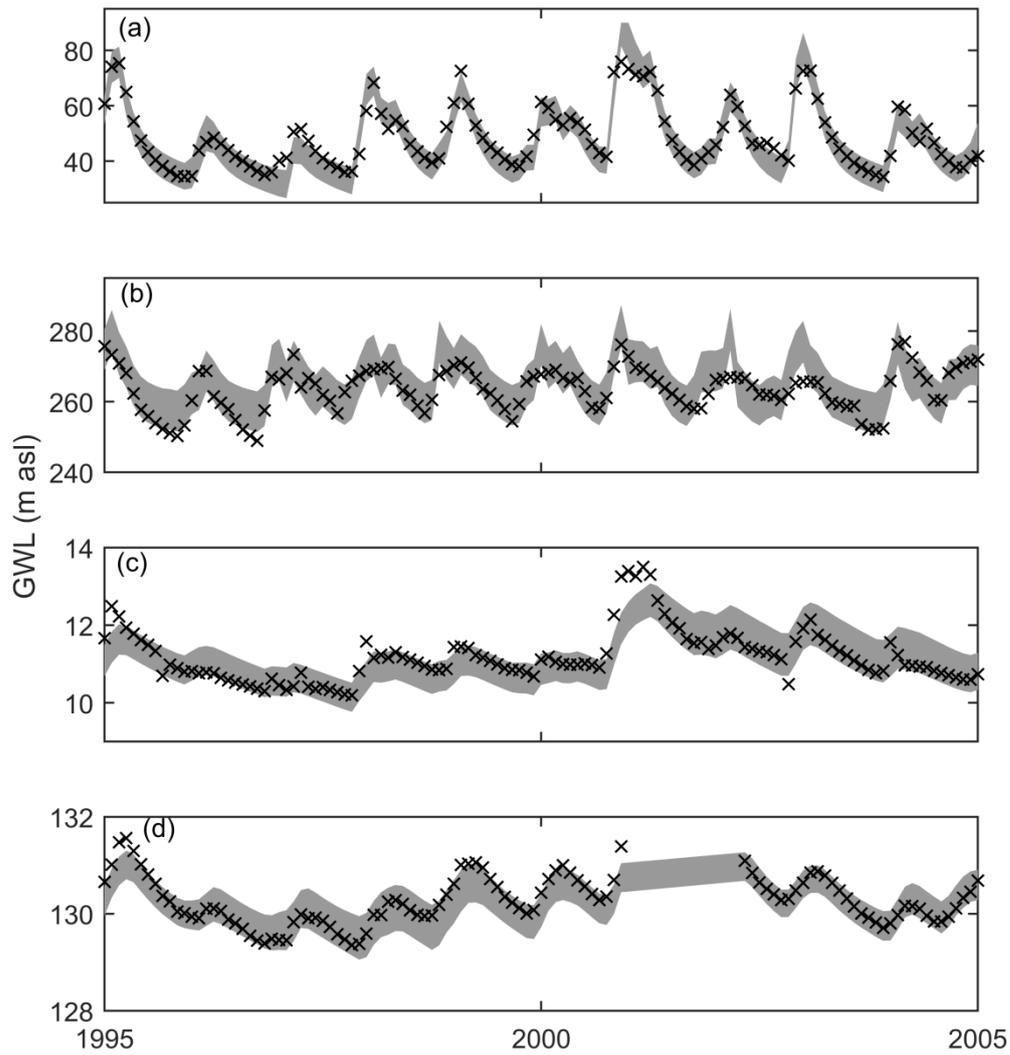
676 **Figure 4:** Maximum likelihood estimate of AEP pdf of residuals (continuous curve) and observed
 677 distribution of residuals during calibration (crosses) and validation (circles) periods. Boreholes are (a)
 678 Chilgrove House, (b) Hucklow South, (c) Lower Barn Cottage, (d) Skirwith. Plots (e) and (f) correspond
 679 to models at Lower Barn Cottage and Skirwith that have been calibrated on all of the available data.



680

681 **Figure 5:** Maximum likelihood estimate of auto-correlation function for residuals (black line) and
 682 90% confidence interval of the correlation function according to the MCMC sample (shaded region).
 683 Boreholes are (a) Chilgrove House, (b) Hucklow South, (c) Lower Barn Cottage, (d) Skirwith. Plots (e)
 684 and (f) correspond to models at Lower Barn Cottage and Skirwith that have been calibrated on all of
 685 the available data.

686



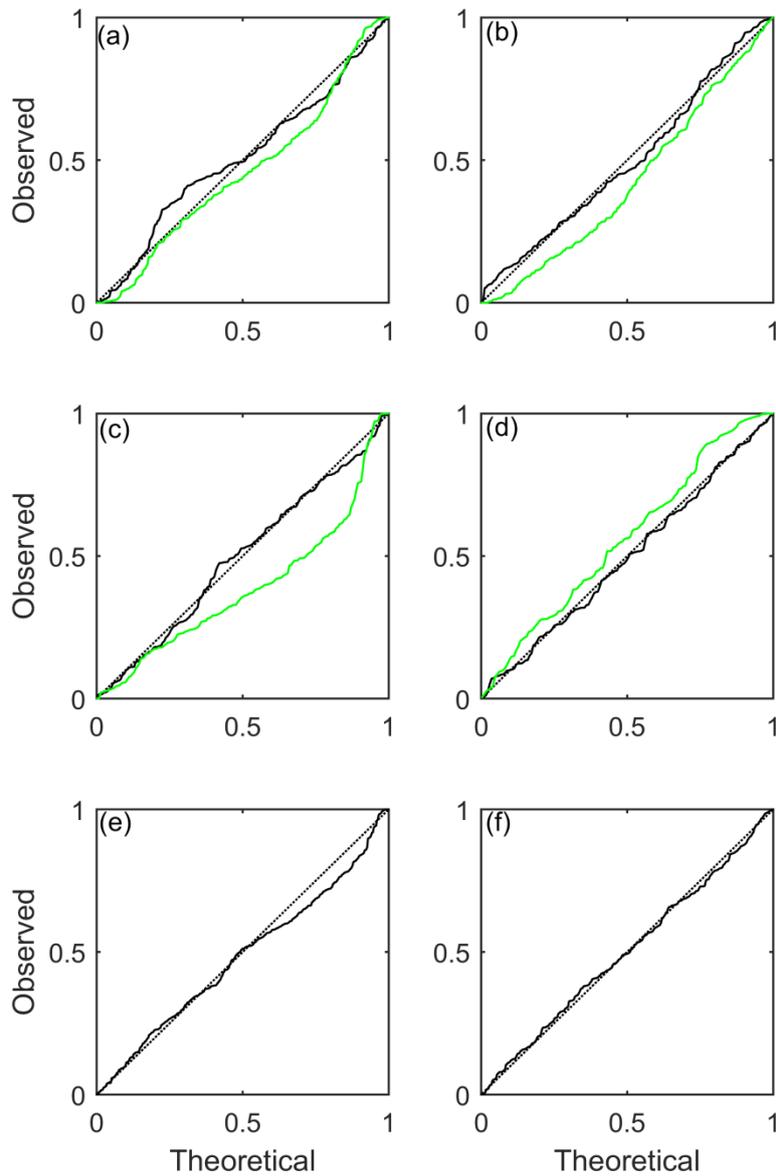
687 .

688 **Figure 6:** 90% prediction intervals of GWLs during a 10-year part of the validation period (shaded
 689 region) and observed GWLs (crosses) according to the best fitting MM. Prediction intervals are based
 690 on the MCMC samples and do not use conditioning data. Boreholes are (a) Chilgrove House, (b)
 691 Hucklow South, (c) Lower Barn Cottage, (d) Skirwith.

692

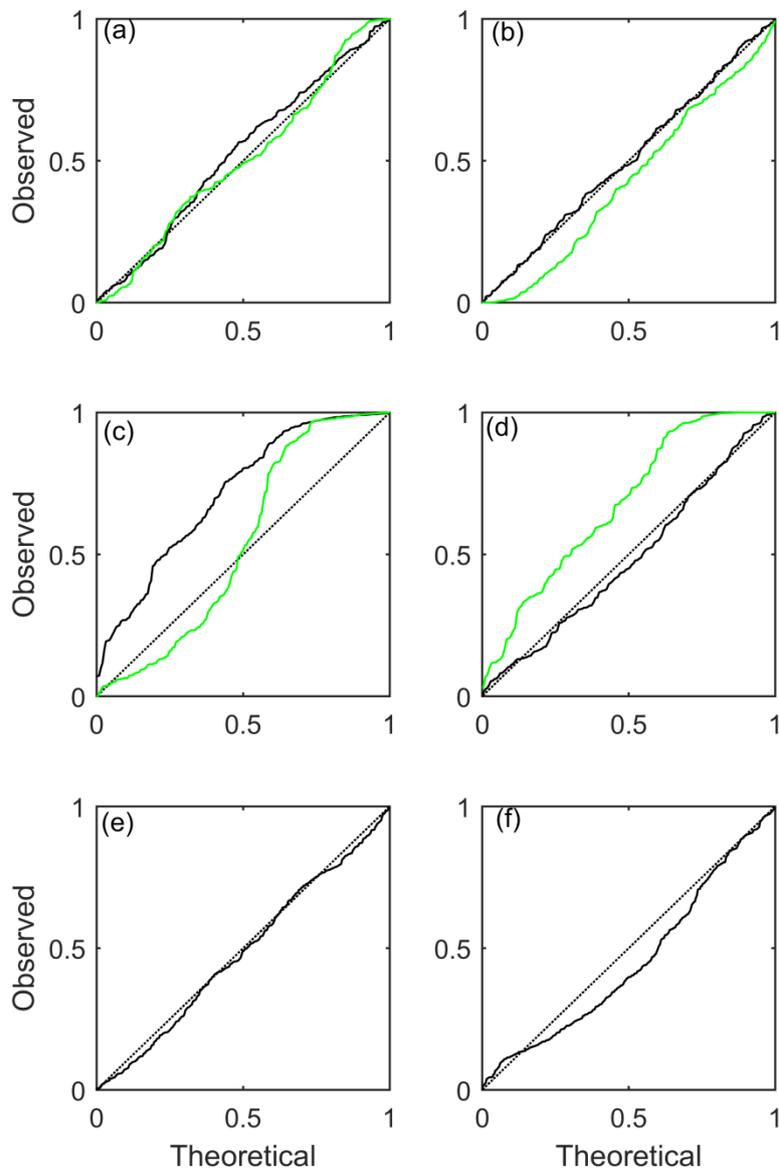
693

694



695

696 **Figure 7:** QQ plots upon prediction of GWLs using maximum likelihood estimate of mixed model with
 697 Gaussian and independent random effects during calibration period (black line) and validation
 698 period (green line). Boreholes are (a) Chilgrove House, (b) Hucklow South, (c) Lower Barn Cottage,
 699 (d) Skirwith. Plots (e) and (f) correspond to models at Lower Barn Cottage and Skirwith that have
 700 been calibrated on all of the available data.

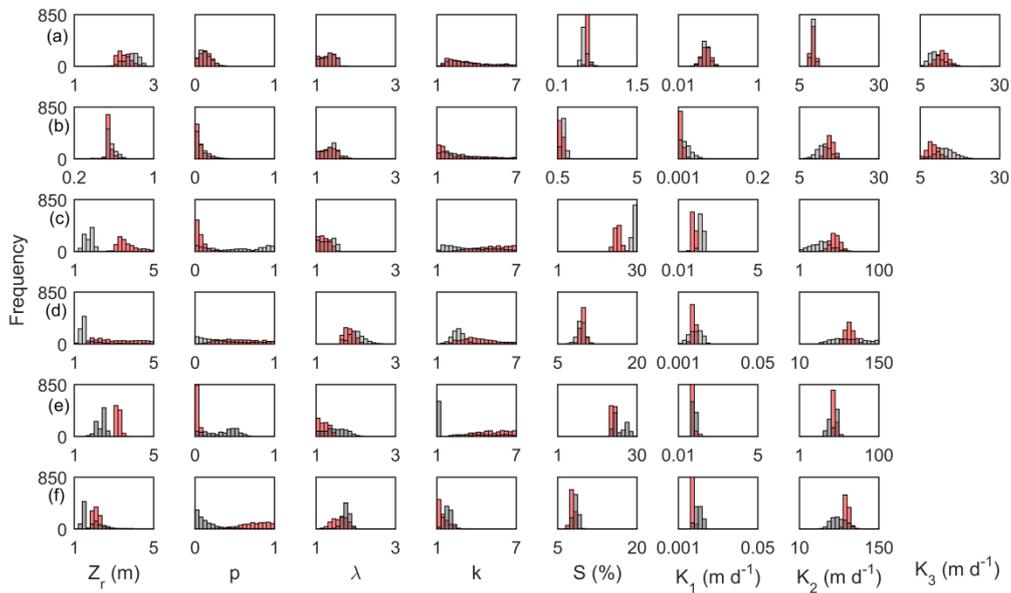


701

702 **Figure 8:** QQ plots upon prediction of GWLs using maximum likelihood estimate of best fitting
 703 generalized mixed model during calibration period (black line) and validation period (green line).
 704 Boreholes are (a) Chilgrove House, (b) Hucklow South, (c) Lower Barn Cottage, (d) Skirwith. Plots (e)
 705 and (f) correspond to models at Lower Barn Cottage and Skirwith that have been calibrated on all of
 706 the available data.

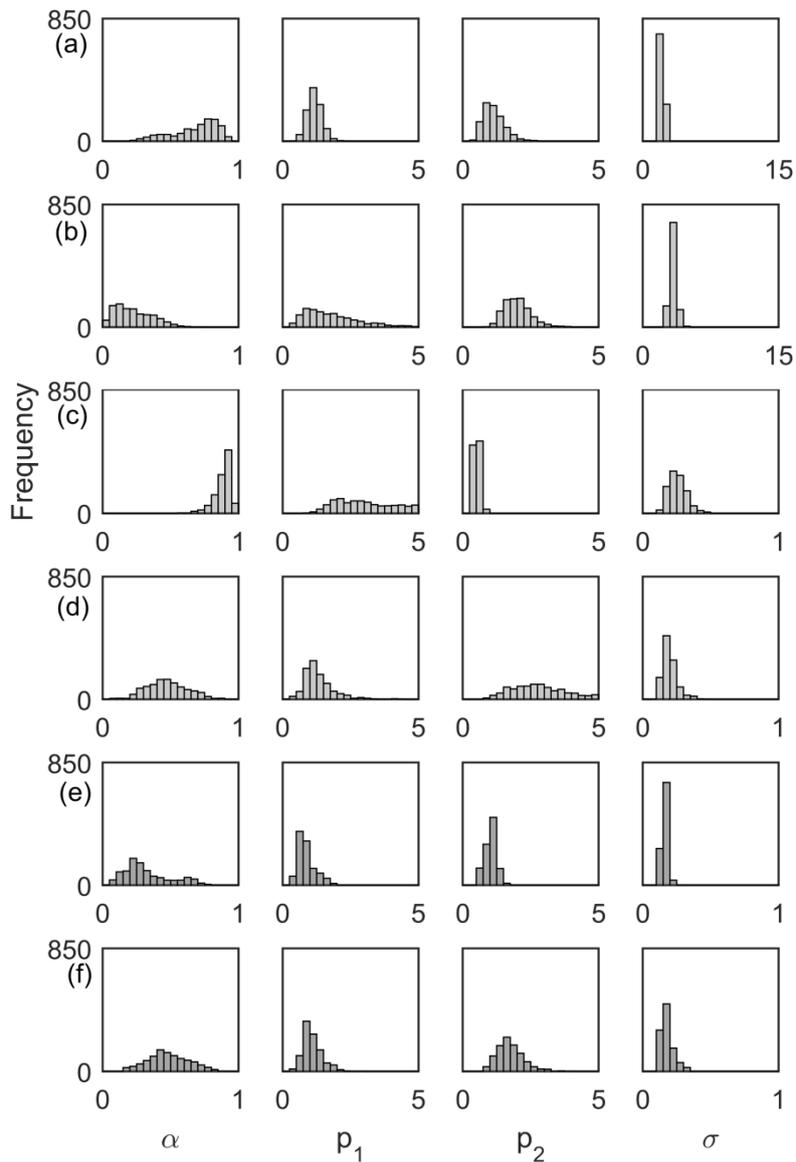
707

708



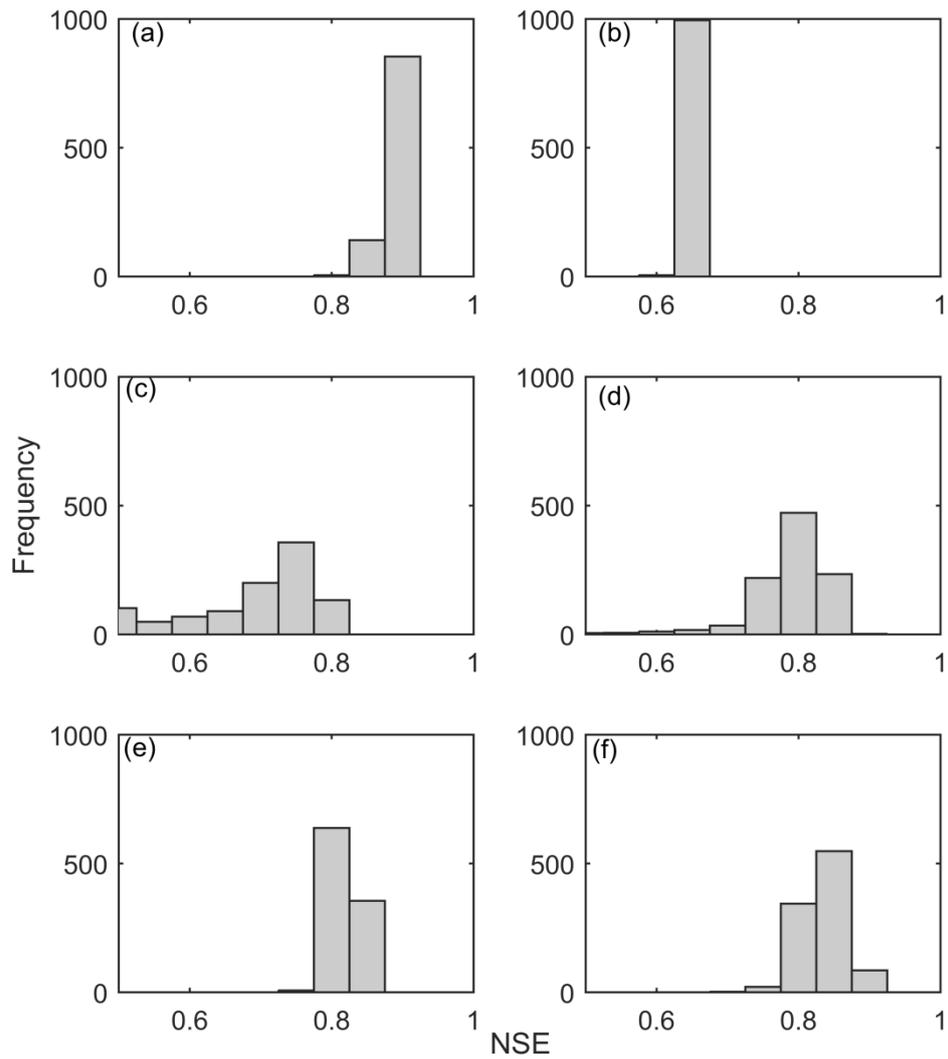
709

710 **Figure 9:** Histograms of Aquimod parameters realized within the 1000 MCMC samples for models
711 with Gaussian independent (red) and AEP generalized (grey) random effects. The bounds on the
712 parameter values correspond to the bounds on the uniform prior distributions. The boreholes are (a)
713 Chilgrove House, (b) Hucklow South, (c) Lower Barn Cottage, (d) Skirwith. Plots (e) and (f) correspond
714 to models at Lower Barn Cottage and Skirwith that have been calibrated on all of the available data.



715

716 **Figure 10:** Histograms of AEP marginal distribution parameters realized within the 1000 MCMC
 717 samples. The bounds on the parameter values correspond to the bounds on the uniform prior
 718 distributions. The boreholes are (a) Chilgrove House, (b) Hucklow South, (c) Lower Barn Cottage, (d)
 719 Skirwith. Plots (e) and (f) correspond to models at Lower Barn Cottage and Skirwith that have been
 720 calibrated on all of the available data.



721

722 **Figure 11:** Calibration NSE scores for the MCMC samples of AquiferMod parameters. The boreholes are

723 (a) Chilgrove House, (b) Hucklow South, (c) Lower Barn Cottage, (d) Skirwith. Plots (e) and (f)

724 correspond to models at Lower Barn Cottage and Skirwith that have been calibrated on all of the

725 available data.

726

727

728

	Gaussian			AEP		
	Independent	Exponential	Matérn	Independent	Exponential	Matérn
Chilgrove House	1247.2	1241.5	1243.1	1219.3	1209.1	1211.1
Hucklow South	1137.1	1133.3	1135.5	1122.5	1113.9	1117.1
Lower Barn Cottage	-87.1	-237.0	-243.8	-108.4	-294.1	-294.6
Skirwith	-162.1	-277.7	-275.4	-155.6	-283.9	-280.8

730

731 **Table 1:** AIC values for maximum likelihood estimates of mixed models for the calibration data from

732 four boreholes with different distributions and correlation functions. Smallest AIC values are

733 highlighted in bold.

734

	Gaussian			AEP		
	Independent	Exponential	Matérn	Independent	Exponential	Matérn
Chilgrove House	0.93	0.93	0.93	0.93	0.93	0.90
Hucklow South	0.74	0.73	0.73	0.73	0.73	0.70
Lower Barn Cottage	0.74	0.65	0.64	0.76	0.76	0.54
Skirwith	0.84	0.82	0.83	0.84	0.84	0.78

735

736 **Table 2:** NSE scores for maximum likelihood estimates of mixed models for the calibration data from

737 the four boreholes with different distributions and correlation functions.