Hydrogeology Journal (2022) 30:1801–1817
https://doi.org/10.1007/s10040-022-02528-y

PAPER

Temporal interpolation of groundwater level hydrographs for regional drought analysis using mixed models

B. P. Marchant[1] · D. Cuba[1,2] · B. Brauns[1] · J. P. Bloomfield[3]

Received: 14 January 2022 / Accepted: 21 July 2022 / Published online: 25 August 2022

Abstract

Large-scale studies of the spatial and temporal variation of groundwater drought status require complete inventories of groundwater levels on regular time steps from many sites so that a standardised drought index can be calculated for each site. However, groundwater levels are often measured sporadically, and inventories include missing or erroneous data. A flexible and efficient modelling framework is developed to fill gaps and regularise data in such inventories. It uses linear mixed models to account for seasonal variation, long-term trends and responses to precipitation and temperature over different temporal scales. The only data required to estimate the models are the groundwater level measurements and freely available gridded weather products. The contribution of each of the four types of trends at a site can be determined and thus the causes of temporal variation of groundwater levels can be interpreted. Validation reveals that the models explain a substantial proportion of groundwater level variation and that the uncertainty of the predictions is accurately quantified. The computation for each site takes less than 130 s and requires little supervision. Hence, the approach is suitable to be upscaled to represent the variation of groundwater levels in large datasets consisting of thousands of boreholes.

Keywords Groundwater monitoring · Statistical modelling · Mixed model · Groundwater drought

Introduction

The groundwater yield from boreholes is typically a function of groundwater level (GWLs; Gleeson and Ingebritsen 2016; Ascott et al. 2019). Groundwater hydrographs provide information about the continuously changing status of groundwater resources at a site. In this context, groundwater hydrographs are useful in the quantification and management of the response of groundwater to meteorological drought (Van Loon 2015). In particular, standardised groundwater hydrographs are used in studies of episodes of major drought to compare differences in the response of groundwater systems between sites (Bloomfield et al. 2015; Marchant and Bloomfield 2018) and with standardised data for other components of the hydrological cycle and driving meteorology to understand the propagation of droughts through the terrestrial water cycle (Folland et al. 2015; Van Loon 2015).

Standardised meteorological and hydrological time series, such as the Standardised Precipitation Index (SPI; McKee et al. 1993), the Standardised Precipitation Evapotranspiration Index (SPEI; Vicente-Serrano et al. 2010), the Standardised Streamflow Index (SSI; Barker et al. 2016; Svensson et al. 2016) and the Standardised Groundwater level Index (SGI; Bloomfield and Marchant 2013), are estimated using a wide range of techniques that include distribution fitting and nonparametric methods (Svensson et al. 2016). These methods all require data on a common, regular time step. Unfortunately, even when GWL data are collected at a nominally consistent frequency, temporal irregularity and missing observations are still commonplace (Marchant and Bloomfield 2018; Peterson and Western 2018).

An additional challenge when working with raw GWL observations, and one particularly pertinent to the study of hydrological extremes such as groundwater droughts, is the presence of outliers. These observations must be analysed in detail and, if they are thought to be erroneous, removed from the data prior to model estimation. Analyses of major groundwater droughts are increasingly making use of

✉ B. P. Marchant
benmarch@bgs.ac.uk

[1] British Geological Survey, Keyworth NG12 5GG, UK

[2] School of Mathematics and Statistics, University of Glasgow, Glasgow, UK

[3] British Geological Survey, Wallingford OX10 8BB, UK

improved access to time series data from hundreds to thousands of observations wells (Kumar et al. 2016; Marchant and Bloomfield 2018), so a final requirement is that any approaches to GWL interpolation and outlier identification and removal should ideally be applicable to large observational datasets and large-sample hydrology problems (Gupta et al. 2014).

Similar challenges exist in the wider environmental sciences such as meteorology and water quality monitoring. Shabalala et al. (2019) and Zhang and Thorburn (2022) review data infilling methodologies in these contexts.

Here, a flexible mixed model (Dobson 1990) approach is described that combines monthly GWL interpolation to infill missing observations with identification and removal of outliers. The method is illustrated with six GWL hydrographs from various aquifers in the UK (Allen et al. 1997). However it is explicitly designed to apply to large-sample problems, such as regional- to continental-scale groundwater drought analysis, where metadata associated with sites may be minimal and may be effectively restricted to gridded meteorological data (Brauns et al. 2020).

Previously applied approaches to temporal interpolation of GWLs differ with regards to (1) whether and in how much detail the processes that drive GWL variation are accounted for and (2) the heuristic or statistical methods used to predict GWLs on dates when they have not been measured. According to anecdotal evidence from Peterson and Western (2018), a variety of heuristic approaches are often adopted—for example, where GWLs are interpolated within a time series between observations to the date of interest; by the adoption of data from temporally closest points; or, by averaging data over some appropriate period. These methods are not ideal since they are generally not easily replicated across sites or between studies and can lack rigorous justification. In contrast, Zaghiyan et al. (2021) consider formal statistical approaches when infilling groundwater data. These approaches follow standard protocols and are more easily replicated but they do not utilise knowledge of the local weather, hydrology and hydrogeology. Marchant and Bloomfield (2018) address this issue using linear mixed models. These models combine linear relationships between GWLs and drivers of variation such as seasonality and precipitation with statistical models of the degree of temporal auto-correlation in the observed data from a site. Other modelling approaches represent the drivers of groundwater variation in more detail. For example, lumped conceptual models for the simulation of GWL time series (Birtles and Reeves 1977; Keating 1982; Kazumba et al. 2008; Mackay et al. 2014) include the movement of precipitation through the hydrogeological system to the location of an individual borehole, whereas physically based process-driven-regional-groundwater models investigate complex groundwater flow systems and the impacts of environmental changes on those systems more widely (Zhou and Li 2011). The linear, conceptual and regional groundwater models can all be used to model GWLs on a regular time step at a location. The more detailed models can lead to more accurate predictions if sufficient data are available for their calibration. However, this is often not the case (Trichakis et al. 2017; Hellwig et al. 2020), and regional groundwater models, in particular, can suffer from challenges associated with equifinality (Beven 2006), where there is insufficient data to assess which physical processes are driving GWL variation.

When modelling the spatio-temporal status of groundwater droughts across 948 GWL monitoring sites in the Chalk aquifer of the UK, Marchant and Bloomfield (2018) chose to adopt a statistical mixed-model approach. They noted that although lumped models are much less demanding of data for calibration than physically based distributed groundwater models, they still need some assumptions to be made regarding the range of possible model structures and may require more parameters to be fitted than equivalent statistical models such as a mixed model. Simultaneously and independently of Marchant and Bloomfield (2018), Peterson and Western (2018) published a model for the temporal interpolation of groundwater hydrographs at a site that combined a soil moisture/recharge model with a statistical model. The approaches of both Marchant and Bloomfield (2018) and Peterson and Western (2018) rely on a combination of an impulse response function (IRF) or linear transfer function noise (TFN) model to account for meteorological forcing of GWLs (von Asmuth et al. 2002; von Asmuth and Bierkens 2005), and some form of temporal interpolation of the portion of groundwater variation that is not explained by this forcing model.

Peterson and Western (2018) used a TFN model that simulated changes in groundwater head by water leaving a soil-water store on a daily time step using daily precipitation and areal potential evapotranspiration (Peterson and Western 2014) applied to a simple soil-moisture partitioning model, whereas Marchant and Bloomfield (2018) estimated monthly changes in head by implementing the precipitation IRF of von Asmuth et al. (2002) combined with sinusoids of periods of 6 and 12 months added as fixed effects to their mixed model to represent seasonal variation caused by evapotranspiration. This latter approach was taken as they did not have appropriate data to calibrate soil-moisture models at their 948 study sites. Peterson and Western (2018) and Marchant and Bloomfield (2018) then both predicted heads at the desired time points using interpolation techniques that are more commonly applied in a spatial context (Webster and Oliver 2007). Due to the simple representation of change in groundwater head in Marchant and Bloomfield (2018), it is more amenable for use in large-sample studies of groundwater drought where supporting metadata related to aquifer properties is usually lacking; however, it has limitations. The model does not explicitly take account of temperature and had limited success in fitting data from sites where there were apparent trends, perhaps affected by extrinsic, anthropogenic factors. The fixed effects component of the mixed model of Marchant and Bloomfield (2018) has been extended here to include further covariates and explicitly account for temperature. Also, the estimation of the mixed models proposed by Marchant and Bloomfield (2018) is time-consuming, and it is impractical to scale up the process to multiple thousands of sites; therefore, efficiencies in the model estimation procedure have been sought.

To date, there have been few studies of groundwater outliers, despite their common occurrence in observational data. Errors may arise due to mistakes in data recording and transcription, but outliers might also reflect changes in the borehole, such as collapse or changes in reference datum or logger position. Data quality 'flags' can be attached to observations during data management; however, systematic identification of outliers is rare. Following a series of data control steps, such as checking well location and that reported groundwater heads were below ground level for unconfined aquifers, Tremblay et al. (2015) identified outliers as those beyond an arbitrary depth threshold related to a local mean level. Li et al. (2016) identified observations of more than three standard deviations from a smoothed hydrograph. Peterson et al. (2018) fitted a double exponential smoothing time-series model to identify a 'noise envelope' and excluded GWL observations outside the envelope. Marchant and Bloomfield (2018) performed cross-validation on their fitted mixed model of GWLs and removed observations with standardised squared prediction errors (SSPE) of greater than 50 before refitting the mixed model. A similar approach to Marchant and Bloomfield (2018) is adopted here in the revised mixed model.

Following a brief description of the case study sites and data, the methodology for estimating GWLs using mixed models with an extended set of covariates is presented along with a description of the outlier identification method. The results for six sites from the UK are presented and discussed in the context of the application of the approach to large-sample analysis of groundwater hydrographs and particularly in the context of regional- to continental-scale drought analysis. The six sites are chosen because of their contrasting patterns of variation with different degrees of seasonality and evidence of long-term trends, and the contrasting number of missing observations. Extensive validation is performed to assess the accuracy of the approach for predicting GWLs on dates when they were not observed, determining the uncertainty of these predictions and estimating SGI values. In the validation exercise, temporal gaps of different lengths are introduced to the observation record to assess the minimum data requirements for its application.

Statistical theory

Linear mixed models

Our temporal modelling procedure is based upon linear mixed models of the form:

$$\mathbf{z} = \mathbf{M}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where $\mathbf{z} = (z_1, z_2, \ldots, z_n)^{\mathrm{T}}$ is a vector of n GWL measurements, and $z_i = z(m_i)$, where m_i is the number of months since the start of the study period; $\mathbf{M}$ is an $n \times q$ design matrix containing q temporally varying covariates recorded at each of the n observation times; $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_q)^{\mathrm{T}}$ is a vector of q regression coefficients, and $\boldsymbol{\varepsilon}$ is a vector containing n normally distributed residuals. The elements of the $\mathbf{M}\boldsymbol{\beta}$ term relate the variation in GWLs to available covariates or driving variables and are referred to as the fixed effects. The $\boldsymbol{\varepsilon}$ residuals or differences between the observed GWLs and the fixed effects are referred to as the random effects. The random effects are assumed to have been realised from a Normal distribution, and have zero mean and covariance matrix $\mathbf{C}$. Nondiagonal elements of $\mathbf{C}$ can be nonzero, which indicates that the random effects can be temporally correlated. This formulation of the model uses data on a monthly time step, although the approach can also be applied to daily recorded data.

The elements of $\mathbf{C}$ can be determined from a parametric covariance function, $C(\tau)$, such as the nested nugget and Matérn function which relates the degree of correlation between a pair of observed groundwater measurements to τ, the number of months separating the measurements:

$$C(\tau) = \begin{cases} c_0 + c_1 \text{ if } \tau = 0 \\ c_1 G(\tau) \text{ for } \tau > 0 \end{cases} \tag{2}$$

where

$$G(\tau) = \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\frac{2\sqrt{\nu\tau}}{a}\right)^{\nu} K_{\nu}\left(\frac{2\sqrt{\nu\tau}}{a}\right) \tag{3}$$

Γ is the gamma function and K_ν denotes the modified Bessel function of order ν. The covariance function has four parameters, the nugget variance c_0, the partial sill variance c_1, a temporal parameter a and a smoothness parameter ν. The variance of the random effects at each time is equal to the sill variance or $c_0 + c_1$. The covariance between two random effects separated by an infinitesimal temporal lag is c_1. The covariance function decays towards zero as τ increases. The smoothness parameter controls the shape of the function for small τ, whereas the temporal parameter controls the timescales over which the random effects are correlated.

In contrast to many time series analysis methods, the linear mixed model does not require that the GWLs are observed at a regular frequency. Indeed, linear mixed models are commonly applied to spatial problems in which measurements are made irregularly across a study area. The covariance matrix can be calculated for observations made at any set of times, and the model estimation procedure accounts for the varying degree of correlation between pairs of observations.

The covariance parameters $\boldsymbol{\alpha}=(c_0,c_1,a,\nu)^{\mathrm{T}}$ and the regression parameters $\boldsymbol{\beta}=(\beta_1,\beta_2,\ldots,\beta_q)^{\mathrm{T}}$ of the linear mixed model must be estimated from available data. This can be achieved with the maximum likelihood estimator which uses a numerical optimization algorithm to find the parameter values which lead to the maximum achievable value of the log-likelihood function:

$$L(\mathbf{z}|\boldsymbol{\alpha},\boldsymbol{\beta})=\text{Constant}-\frac{1}{2}\ln|\mathbf{C}(\boldsymbol{\alpha})|-\frac{1}{2}(\mathbf{z}-\mathbf{M}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{C}(\boldsymbol{\alpha})^{-1}(\mathbf{z}-\mathbf{M}\boldsymbol{\beta}) \tag{4}$$

In general, this optimisation is computationally demanding. The log-likelihood must be calculated for sufficient sets of parameter values to determine the values of the $q+4$ parameters that result in the largest likelihood. Each calculation of the likelihood function requires the computation of the Matérn covariance function for each element of the covariance matrix and then the inversion of this matrix. The computational expense of the optimisation can be reduced by differentiating Eq. (4) and noting that for a given $\boldsymbol{\alpha}$ the log-likelihood is maximised when

$$\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}=(\mathbf{M}^{\mathrm{T}}\mathbf{C}(\boldsymbol{\alpha})^{-1}\mathbf{M})^{-1}\mathbf{M}^{\mathrm{T}}\mathbf{C}(\boldsymbol{\alpha})^{-1}\mathbf{z} \tag{5}$$

Further efficiencies can be achieved by constraining the parameters of the linear mixed model. For example, the smoothness parameter, ν, of the Matérn function could be fixed rather than estimated. In, particular, if ν is set to 0.5 then the nested nugget and Matérn function reduce to the nested nugget and exponential model:

$$C(\tau)=\begin{cases} c_0+c_1 & \text{if } \tau=0 \\ c_1\exp\left(-\frac{\tau}{a}\right) & \text{for } \tau>0 \end{cases} \tag{6}$$

This leads to great computational efficiencies because each evaluation of the covariance matrix no longer requires calculation of gamma and Bessel functions and because fewer evaluations of the log-likelihood function are required to estimate one fewer parameter.

Variable selection

When using linear mixed models, a practitioner must decide which covariates to include in the fixed effects. The fixed effects should reflect the proportion of the variation of GWLs that can be explained by seasonal variation, weather variables or long-term trends. Omission of key covariates could mean that the fixed effects explain a smaller than necessary proportion of the variation. If too many covariates are included, the model might be overfitted, meaning the model is too closely tuned to features of the calibration data so it predicts the GWL at other times relatively poorly. Various statistical tests can be used to assess whether the inclusion of a particular covariate is justified by the resultant improvement in the model fit. The Akaike Information Criterion (AIC; Akaike 1973):

$$\mathrm{AIC}=2k-2L \tag{7}$$

where k is the number of estimated model parameters, is often used to compare linear mixed models of spatial variation (Webster and Oliver 2007). The set of covariates that minimises the AIC is thought to best manage the trade-off between model complexity (number of parameters) and the model fit (the likelihood).

The result of any assessment of the necessity of including a particular covariate in a model might vary according to which covariates are also present. Hence, the order in which covariates are introduced to the model could influence which covariates are eventually selected—for example, increased water abstraction from an aquifer could lead to a steady decrease in the GWLs. This could be explained in a linear mixed model by including the rate of abstraction as a covariate in the fixed effects. Alternatively, the temporal trends that result from abstraction could be represented by including the date of observation as a covariate. In isolation, either of these covariates is likely to improve the model, but if the other covariate is already present, they might not lead to a substantial further improvement. It is therefore important to preselect covariates that could potentially be included in the linear mixed model to ensure that they are not strongly correlated and likely to explain similar patterns of variation.

Stepwise regression procedures are often used to manage the order in which the potential covariates are considered. A stepwise procedure using forward selection starts from a linear mixed model where the fixed effects are constant. Then a series of models are estimated where the fixed effects are a constant and one of the potential covariates that are being considered. The covariate that leads to the smallest AIC is added to the model provided that this AIC value is less than that for the model with constant fixed effects. The procedure then continues until the addition of any of the remaining covariates fails to improve the AIC value. A backwards-selection procedure commences with a model that includes all of the potential covariates and then iteratively removes covariates until the removal of any of the covariates causes the AIC to increase. Such stepwise methods can be computationally demanding since they require the estimation of multiple linear mixed models.

Prediction

Once the parameters of a linear mixed model have been estimated, it can be used to predict the GWL at any time where it has not been observed using the empirical best linear unbiased predictor (E-BLUP; Lark et al. 2006). The predictor combines the values of covariates at the prediction time, and the regression relationship within the fixed effects with the GWLs observed close to the prediction time. The covariates included in the fixed effects design matrix must be known for all times where predictions are required. The relative weight given to the regression relationship and the observations will depend on the degree of temporal correlation within the random effects and the time lag between the available observations and the prediction time. If the random effects are independent, the predictor only uses the regression relationship. For the linear mixed model with Normal fixed effects (Eq. 1) the E-BLUP outputs are the expected GWL, $\hat{z}_i$, and the variance of this prediction, σ_i^2 after m_i months. These outputs are sufficient to specify the entire probability density function (pdf) of the groundwater prediction m_i months after the start of the study period and to determine the probability that any specified GWL might have occurred.

Validation

The use of a specific linear mixed model implies various assumptions about the variation of GWLs. Validation of these assumptions is crucial to ensure the accuracy of the predictions. Also, validation can identify problems in the model estimation procedures and indicate outliers amongst the measurements which might be erroneous.

The model should accurately predict GWLs without bias and the predicted model variances should reflect the size of the errors in the model predictions. Validation procedures compare model predictions to observed data. Ideally, the data would not have been used to calibrate the model but a sparsity of data or the time required to repeatedly calibrate a model with different subsets of the data can make this impractical. In this circumstance, a ten-fold cross-validation might be performed. Each of the observations are randomly allocated to one of ten 'folds'. The linear mixed model is calibrated using all of the available data but at the prediction stage, one fold is omitted and the model is used to predict the GWL at the times of the observations within that fold. The prediction process is repeated missing out a fold in turn so that each observation can be compared to a predicted value.

The metrics used to assess model accuracy when validation has occurred at n_{v} times include the mean error:

$$\mathrm{ME} = \frac{1}{n_{\mathrm{v}}} \sum_{i=1}^{n_v} \left\{ z_i - \hat{z}_i \right\} \tag{8}$$

and the root mean squared error (RMSE),

$$\mathrm{RMSE} = \left[\frac{1}{n_{\mathrm{v}}} \sum_{i=1}^{n_v} \left\{ z_i - \hat{z}_i \right\}^2 \right]^{\frac{1}{2}} \tag{9}$$

Unbiased and accurate predictions will lead to small values of these metrics but Eqs. (8) and (9) cannot be used to compare the effectiveness of modelling at different locations where the groundwater measurements have different variability. If such comparisons are relevant, then the proportion of variance explained by the model or the correlation between the observed GWLs and the modelled values might be calculated. The Pearson correlation coefficient indicates whether a linear transformation of values of a variable can lead them to be similar to values of another variable. Observations and model predictions should be similar without applying a linear transformation. Therefore, Lin's concordance coefficient is a more relevant validation metric:

$$\rho_{\mathrm{c}}(\mathbf{x}, \mathbf{y}) = \frac{2\rho(\mathbf{x}, \mathbf{y}) \sqrt{\mathrm{var}\,(\mathbf{x})} \sqrt{\mathrm{var}\,(\mathbf{y})}}{\mathrm{var}\,(\mathbf{x}) + \mathrm{var}\,(\mathbf{y}) + \left(\mu_x - \mu_y \right)^2} \tag{10}$$

where $\rho_{\mathrm{c}}(\mathbf{x}, \mathbf{y})$ is Lin's concordance coefficient for variables $\mathbf{x}$ and $\mathbf{y}$, $\rho(\mathbf{x}, \mathbf{y})$ is Pearson's coefficient, var($\mathbf{x}$) is the variance of $\mathbf{x}$ and μ_x is the mean of $\mathbf{x}$. Lin's concordance coefficient can take values between –1 and 1, and a value of 1 indicates an exact match between the two variables.

The prediction variances or uncertainty of the predictions can be assessed by calculating the standardised prediction error (SPE) at each validation time:

$$\theta_i = \frac{z_i - \hat{z}_i}{\sigma_i} \tag{11}$$

If the linear model has correctly quantified the uncertainty, then the θ_i should be realisations from a standardised (zero mean and unit variance) Normal distribution, and the mean squared SPE should be 1 (Lark et al. 2006). The θ_i can also be used to assess whether the model indicates that an observation is very unlikely to occur and is an outlier—for example, there is a probability of less than 0.003, according to the model, of the observed value leading to a magnitude of θ_i greater than 3.

Methods

Case study sites and data

Groundwater level data from six UK boreholes were considered. Three of the sites (A–C in Fig. 1) are located on the Permo-Triassic sandstone aquifer, and three on the Cretaceous Chalk

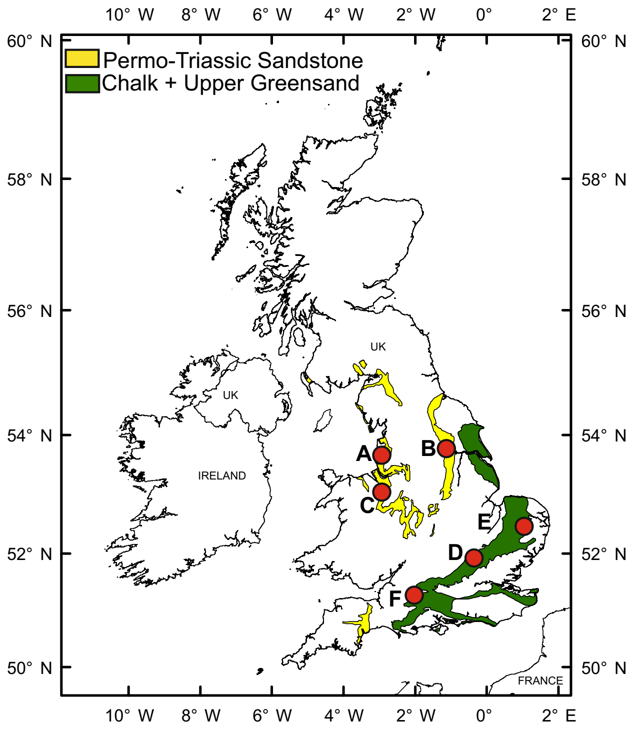

Fig. 1 Locations of the six study boreholes in the UK, © UKRI 2022

aquifer (D–F). The Chalk and Permo-Triassic sandstone aquifers (Fig. 1) are the two most important aquifers in the UK each providing regionally important public water supplies, water for agriculture and industry, and in the case of the Chalk, significant baseflow to overlying rivers and wetlands (Allen et al. 1997; Jones et al. 2000; Bloomfield et al. 2009). All six sites are observation, not abstraction, boreholes and part of the national GWL monitoring network. Although Heudorfer et al. (2019) have recently proposed 45 indices to classify groundwater hydrographs based on three principal classes of descriptors (structure, distribution and shape), the six hydrographs used here have been selected to reflect the typical variation in frequency and density of observations as much as the variety of form in GWL hydrographs in the UK. Notwithstanding this, hydrographs were selected (based on visual inspection) that showed both strong seasonality (sites D and F), extra-annual correlation (C and E), and upward (A) and downward (B and C) trends. The monthly medians of the observed data are included in Fig. 2. All six hydrographs have irregular temporal observations and periods of missing data over the analysis period, January 1980 to December 2015—for example, at site C, GWL observations are approximately weekly between 1980 and 1988 and approximately monthly from 1989 to 1994. In 1995 there are only four approximately quarter annual observations and from 1996, the observations are typically daily.

Groundwater level data for the six sites were obtained from the British Geological Survey (BGS) database (British Geological Survey 2020). Table 1 summarises the level data and hydrogeological characteristics of each site. Observation boreholes in the UK are typically open over extended depth intervals representative of the full thickness of the active aquifer. This is the case for the observation boreholes in this study (Table 1), although note that slotted casing is present in the three Permo-Triassic observation boreholes in the upper-most sections of the boreholes to prevent collapse. Consequently, groundwater levels from such observation boreholes reflect piezometric heads in the regional groundwater system. The length of the record, as well as the sampling frequency, varies from borehole to borehole. When multiple measurements were available within a month, the median was calculated. As is often the case, the records were not complete and had a varying number of missing observations. Table 1 includes a note on the total number of 'missing' months for each hydrograph.

Both monthly precipitation and air temperature data are used in the mixed model to estimate monthly GWLs. The precipitation data were downloaded from the Global Precipitation Climatology Centre (GPCC), version 2018 (Schneider et al. 2018). A gridded product, the data has a spatial resolution of 0.25 × 0.25° with a temporal coverage ranging from January 1891 to December 2016. A precipitation time series was extracted for all boreholes by selecting the grid cell closest to the borehole location. For the temperature variable, the dataset selected (CRU TS v 4.03) was sourced from the Climate Research Unit (CRU) based at the University of East Anglia (Harris et al. 2020). The data were interpolated into a 0.5 × 0.5° grid using angular weighted distance. The data cover the period starting from 1901 and ending in 2018 with monthly frequency. Like the precipitation dataset, a temperature time series was obtained for each borehole by extracting the data for the grid cell closest to the borehole location.

Statistical modelling

Linear mixed models

Linear mixed models (Eq. 1) were used to represent the variation of monthly GWLs at each of the six locations considered in this study. The models could potentially be reformulated to consider GWLs with any other temporal frequency. Different covariates were included in the fixed effects to represent (1) a constant, (2) seasonal variation, (3) long-term temporal trends, (4) responses to rainfall and (5) responses to temperature:

1. *A constant.* A nonzero constant can be accommodated in the fixed effects by including a column in the fixed effects design matrix (Eq. 1) where every row is 1.
2. *Seasonal variation.* Sinusoidal terms with period 12 months represented the seasonal variation. Two columns were included in the fixed effects design matrix, equal to:

$$\sin\left(\frac{2\pi m_i}{12}\right) \text{ and } \cos\left(\frac{2\pi m_i}{12}\right) \tag{12}$$

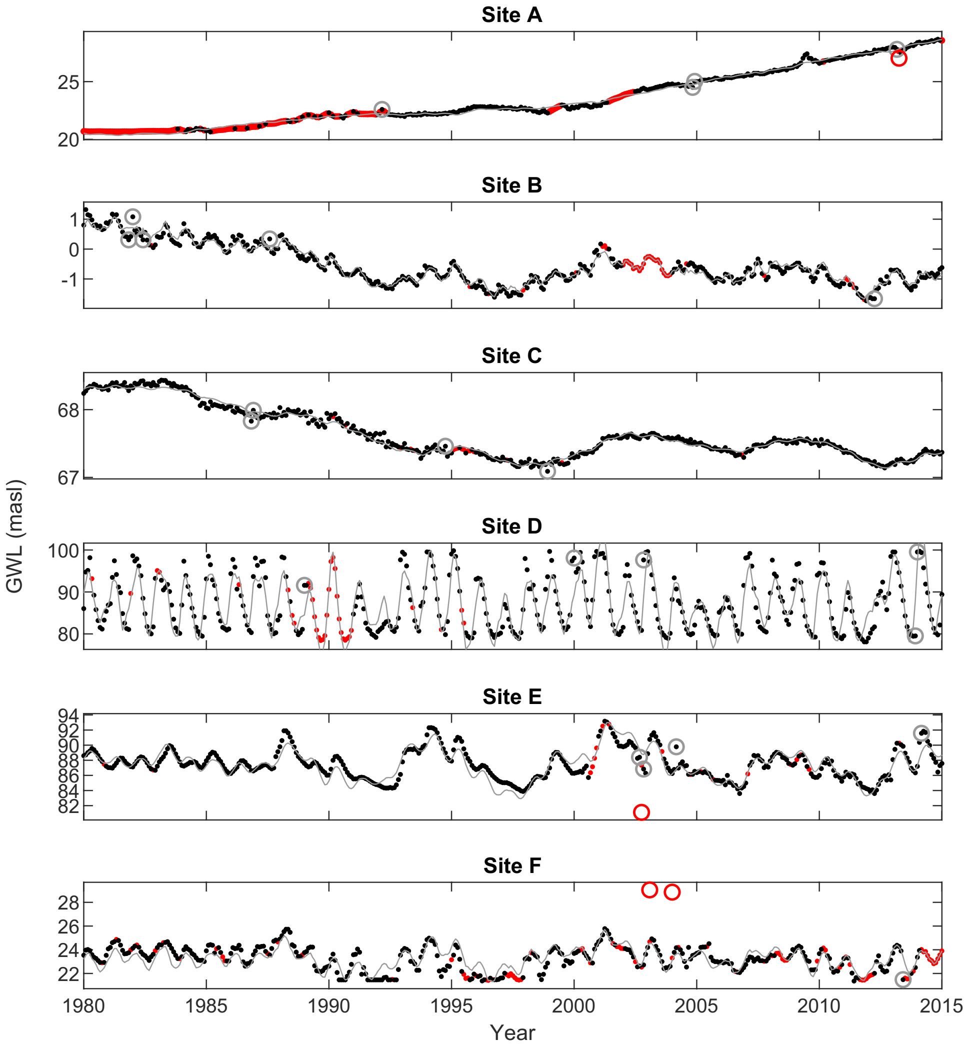

Fig. 2 Groundwater level (GWL) observations and model predictions for the six sites. Observed GWLs (black dots), predicted GWLs for months where observations are missing (red dots), removed outliers (red circle), outliers according to absolute SPE threshold of three (grey circles) predicted fixed effects (grey line)

The estimated regression coefficients controlled both the magnitude of the sinusoid and its phase (i.e. the month of the year in which the seasonal term was largest).

3. *Long-term temporal trends.* A linear trend in GWLs throughout the temporal window of the data can be included in the fixed effects via a column of the design

Table 1 Summary site metadata and statistics (minimum, maximum, mean, standard deviation (s.d.) and number of months with no observations) for GWLs observed at six UK boreholes. *aOD* above Ordnance Datum, *bgl* below ground level, *masl* meters above sea level

Borehole	Datum (m aOD)	Total depth (m)	Completion	Aquifer	Min (m asl)	Max (m asl)	Mean (m asl)	s.d. (m asl)	Missing
A	48.14	100	Casing to 41.0 m bgl	Permo-Triassic Sandstone	20.64	28.64	24.41	2.25	150
B	4.77	45.7	Casing to 20.5 m bgl	Permo-Triassic Sandstone	-1.73	1.32	-0.55	0.66	39
C	74.57	36.6	Casing to 21.9 m bgl	Permo-Triassic Sandstone	67.09	68.43	67.63	0.37	16
D	98.76	21.7	Open hole	Cretaceous Chalk	78.11	99.84	87.19	6.49	34
E	138.17	80.8	Open hole	Cretaceous Chalk	81.11	93.17	87.46	1.95	20
F	29	21.3	Open hole	Cretaceous Chalk	21.36	29.05	23.32	1.09	91

matrix equal to m_i. However, many long-term trends are likely to be nonlinear, perhaps reflecting changes to the amount of water extracted from the aquifer. Nonlinear trends can be accommodated using spline basis functions as columns of the fixed effect design matrix (Marchant 2021). These relatively smooth functions each focus their nonzero values on a different portion of the study period. When multiplied by corresponding regression coefficients, such basis functions can lead to highly flexible and smooth nonlinear trends. The times at which the different spline basis functions meet are called knots. Cyclic spline basis functions (Wood 2017) include a constraint that the values of the spline term at the start and end of the study period are identical. In this paper, four cyclic spline basis functions are combined with a linear trend. The knots for the spline functions are evenly spaced so that each function covers 9 years. The linear trend term controls the difference between the fixed effects at the start and end of the temporal window, whereas the cyclic spline terms control the deviations from the linear trend. If the statistical significance of the linear trend regression coefficient term is tested, then this test assesses whether GWLs have changed during the temporal window (after accounting for the other fixed effects), but this change is not necessarily linear. The mean of fixed effects produced from cyclic spline basis functions can be controlled according to the value of their regression coefficients. Therefore, the constant fixed effect term is removed when the cyclic splines are included in the model.

4. *Responses to precipitation.* According to a standard linear mixed model (Eq. 1), the variable of interest, such as GWLs, immediately respond to changes in the covariates included in the fixed effects design matrix. However, the GWL response to precipitation can occur over timescales of multiple months or years, with the exact timescale varying between sites according to the hydrogeological setting and the time required for water to flow to the groundwater store (Bloomfield and Marchant 2013). Therefore, it is not sufficient to include a time-series of monthly precipitation as a column of the fixed effects design matrix. Instead, it is necessary to account for the precipitation over multiple months prior to each observation. This could be achieved by including n_p precipitation time series $p(m_i-\tau)$ for τ from 0 to $n_\mathrm{p}-1$ as columns of the design matrix, but that would lead to a model with a large number of regression parameters that is likely to have a large AIC value (Eq. 10). Alternatively, the average rainfall over the n_p months could be included as a single column of the fixed effects design matrix. However, this would imply that the rainfall $n_\mathrm{p}-1$ months prior to an observation date controls the GWL to the same degree as the rainfall in months much closer to the observation date. Von Asmuth et al. (2002), and later Marchant and Bloomfield (2018), modelled lagged response to precipitation at a location using an IRF and an observational time series of precipitation. Such a term could be included via a single column of the fixed effects design matrix with rows equal to:

$$\sum_{\tau=0}^{n_\mathrm{p}-1} r_\mathrm{p}(\tau)p\left(m_i-\tau\right) \tag{13}$$

where $r_\mathrm{p}(\tau)$ is the precipitation IRF and $p(m_i)$ is the average precipitation for month m_i from the start of the study period. The IRF can be written

$$r_\mathrm{p}(\tau)=\frac{A_\mathrm{p}{}^{s_\mathrm{p}}\tau^{s_\mathrm{p}-1}\exp\left(-a_\mathrm{p}\tau\right)}{\Gamma\left(s_\mathrm{p}\right)} \tag{14}$$

where A_p, a_p, and s_p are parameters and $\Gamma(s_p)$ is the gamma function of order s_p.

In this paper, the precipitation term is expressed as an IRF in the form of Eq. (14). For each IRF, the value of A is selected to ensure that the $r_\mathrm{p}(\tau)$ in Eq. (14) sums to one for $\tau=0,1,\ldots,n_\mathrm{p}-1$. This permits meaningful visual comparison of the IRFs from different sites but does not affect the model fit since this term is scaled by a regression coefficient in the linear mixed model. The a_p, and s_p parameters are optimised along with the $\boldsymbol{\alpha}$ parameters to maximise the likelihood function (Eq. 4) and the $\boldsymbol{\beta}$ parameters are selected according to Eq. (5). The number of months of precipitation included in the model is set to $n_\mathrm{p}=120$. Since the IRF parameters have to be estimated by the numerical optimiser, the inclusion of such a term adds to the number of computations required to calibrate the model.

The raw driving precipitation time series could contain seasonal variation. This seasonal behaviour was removed prior to including precipitation in the linear mixed model by subtracting the mean precipitation for that calendar month from each observed precipitation value.

5. *Responses to temperature.* Air temperature can influence GWLs since it controls the proportion of precipitation that evaporates before it reaches the groundwater store. Like rainfall, the effects of temperature are unlikely to be instantaneous, and therefore an impulse response approach is applied with a column of the fixed effects design matrix of the form:

$$\sum_{\tau=0}^{n_\mathrm{t}-1} r_\mathrm{t}(\tau)t\left(m_i-\tau\right) \tag{15}$$

where $r_\mathrm{t}(\tau)$ is the temperature IRF and $t(m_i)$ is the average temperature for month m_i from the start of the study period. The temperature IRF is defined according to Eq. (14) with parameters A_t, a_t, and s_t. In this paper $n_t=120$.

Again, seasonality in the temperature time series for each location is removed by subtracting the mean value for that calendar month.

Model estimation and variable selection

The fixed effects terms described in the preceding have been chosen to act largely over different timescales. Therefore, there should not be strong correlations between the columns of the fixed effects design matrix and it should be possible to uniquely determine the β_i. Also, it should be possible to consider the inclusion of the different fixed effects terms separately, or in a prespecified order, without this greatly affecting the terms that are eventually selected.

In this paper, rather than using a computationally expensive stepwise approach to variable selection, they are selected through a series of tests in a prespecified order. In each case, a fixed effect term is included if it leads to a decrease in the AIC. First, a model with constant fixed effects is estimated, then the other fixed effects terms are considered in the following order:

1. The seasonal term
2. The long-term trend
3. The precipitation term
4. The temperature term

These terms were selected because they are likely to account for a large proportion of the variation in GWLs and covariate data and because the mixed effect model terms can be easily formulated. Other processes not included in the fixed effects (e.g. abstraction and hydrogeological setting) will impact the random effects and hence will be accounted for in model predictions, although their impact will not be readily discernible from other drivers of GWL variation.

All computations are carried out using the British Geological Survey's Geostatistical Toolbox for Earth Scientists (Marchant 2018), which can be obtained from the corresponding author. The computation time required for variable selection, estimation of the linear mixed model parameters and prediction of GWLs at unobserved times is recorded.

Prediction and validation

The estimated model for each site is used to predict GWLs for all months where observations were not available using the E-BLUP (Lark et al. 2006). In some circumstances, this might require the predictor to extrapolate beyond the information contained in the available data—for example, if no observations are available for the initial years of the study period, then the estimated regression coefficients for long-term temporal trends are unconstrained by data for these years. Prediction of GWLs in this period can be misleading and unreliable. This is particularly true for the highly flexible cyclic spline functions which can rapidly increase or decrease to implausibly large or small values. In this paper, the long-term temporal trend covariate values for prediction months before the first observation month are replaced by the covariate values for the month of the first observation. If no observations are available, a similar change is also made at the end of the temporal window. This implies that any changes to the long-term temporal trends cease beyond the bounds of the data. Although these alterations remove some implausible trends, the results for months where they have occurred should be treated with caution because there is no evidence to support the assumption that trends cease. Related issues can arise when the random effects are autocorrelated over a long timescale. When there is a large gap between the prediction time and the observation times, the random effects tend toward zero. This can mean that when the random effects include long-term temporal trends and observations are not available at the start or end of the study period, the random effects can decay towards zero beyond the scope of the data leading to unlikely patterns of variation.

Ten-fold cross-validation was performed for all observations at each site and the SPEs (Eq. 14) were calculated. The SPEs are used to indicate potentially erroneous measurements that are inconsistent with the statistical model. The choice of threshold on the SPEs is somewhat subjective. If a magnitude of 3 is selected, then on average, one observation amongst the 432 months of the study period is likely to exceed this threshold purely because of the random variation that is consistent with the model. Also, any statistical model simplifies reality and is not exactly consistent with complex hydrogeological variation. Therefore, additional outliers might be identified which reflect the imperfect fit of the model rather than erroneous measurements. In this paper, thresholds on the SPEs of magnitude 3 and 6 are considered. Once outliers have been identified, they are removed from the observational record and the model is refitted and ten-fold cross-validation is performed again.

The effectiveness of using the E-BLUP to infill GWL time series is explored for different-sized gaps in the data record at each site. This procedure consists of selecting an observation, j, at random where observations are available for months $m_j + \tau$ and $m_j - \tau$ for a specified integer lag τ. All observations between months $m_j + \tau$ and $m_j - \tau$ are removed and the remaining observations and the estimated model are used to predict the GWL for month m_j. The process is repeated 500 times for each τ and the mean error, root mean squared error, mean SSPE and Lin's concordance coefficient are calculated.

The approximate proportion of variation explained by each term in the fixed effects was also calculated. First, the variance of the differences between the observed GWLs and the full modelled fixed effects was determined. Then, this variance was recalculated having removed the term of interest from the fixed effects. The approximate proportion of variance explained by the removed term was assumed equal to the increase in variance upon removal of the term divided by the variance of the observations. Note, that for these approximations of the explained variance, the random effects are ignored.

Standardisation of GWL time series

Once missing observations with the GWL time series for a site have been infilled, the de-seasonalised and normalised SGI for that site is calculated using the nonparametric approach described by Bloomfield and Marchant (2013).

Results

Estimation of linear mixed models

From the initial estimation of the linear mixed models with an exponential covariance function, a total of 27 observations with absolute SPE greater than 3 were identified across the six sites (Fig. 2). If the observations had been realised from the estimated linear mixed models for each site, the expected number of observations to exceed this threshold would have been 7. However, according to visual inspection of Fig. 2, the majority of these 27 observations only appear to be minor deviations from the underlying variation of the time series and cannot be considered to be the result of an obvious error in the measurement or data handling procedures. Four observations exceed the larger threshold of 6 on the absolute SPEs. These observations have more apparent deviations from the underlying variation. The three outlying observations from sites E and F can be considered to be global outliers since they are outside the range of the other observations within these time series. There is also a local outlier amongst the observations from site A which is within the range of values recorded at other times for that site but substantially smaller than the values from neighbouring times. The observations where the larger threshold was exceeded were removed from the data record and the linear mixed models were re-estimated.

Upon re-estimation of the linear mixed models, the seasonal sinusoidal term led to a decrease in the AIC and was included in the fixed effects for all sites except for site A (Table 2). The seasonal term explained less than 10% of the variance at sites B, C, D and E, but a more pronounced seasonal pattern of variation was evident at site D (52.9% of variance) and site F (20.7 % of variance). The long-term trend led to decreases in the AIC for sites A, B and C in accordance with the visually apparent trends. The percentage of variance explained at these sites was 98.9, 60.0 and 98.9, respectively.

The precipitation IRF was included in all six models. For sites A, C, D, E and F, the shape of the IRF showed a relatively rapid increase from lag zero up to a maximum and a more gradual decline whereas the function for site B declines monotonically and slowly from its lag zero value (Fig. 3). The maxima are attained after 4 months for site A, 22 months for site C, 2 months for site D, 8 months for site E and 4 months for site F. The initial increase reflects a delay in the precipitation moving through the hydrological system and reaching the groundwater store, whereas the decline reflects the time that the water remains in the system. The decline is fastest for sites A and D and no influence of precipitation is apparent after a lag of 20 months. At sites E and F, the IRF decays to zero after around 45 and 35 months, respectively, whereas it is greater for more than 60 months at site C.

The temperature IRF leads to a decrease in the AIC for only sites C and D where it explains 0.3 and 2.6% of the variance, respectively. As with the precipitation IRFs for these sites, the temperature function for site D acts over a considerably smaller time scale than that for site C.

The contributions of the long-term, precipitation and temperature fixed effects terms are shown in Fig. 4. These factors make quite distinct and not strongly correlated contributions at each site, suggesting that they can each be estimated separately.

The auto-correlation functions for sites A, B and C (Fig. 5, top row) indicate that the vast majority of variation at these sites is explained by the fixed effects. For sites D, E and F, the random effects have variance equal to 0.21, 0.22 and 0.38 times the variance of the observations, respectively. For site D the auto-correlation is negligible for lags of more than 5 months, whereas for sites E and F, some auto-correlation is apparent beyond lags of 10 months.

The computation time required to select the covariates and estimate the linear mixed models varied between 98 and 130 s for the six sites (Table 3). The variation in these times will have been partly caused by the differing number of observations and partly by the time required by the numerical optimisers to converge to parameter values that minimise the likelihood.

Table 2 Approximate percentage of variance explained by each fixed effects term within the linear mixed models. 'NA' values indicate that the fixed effect term was not included in the model

Borehole	Seasonal	Long-term	Precipitation	Temperature
A	NA	98.9	0.3	NA
B	2.6	60.0	24.8	NA
C	0.1	98.8	11.3	0.3
D	52.9	NA	25.8	2.6
E	9.3	NA	71.2	NA
F	20.7	NA	39.7	NA

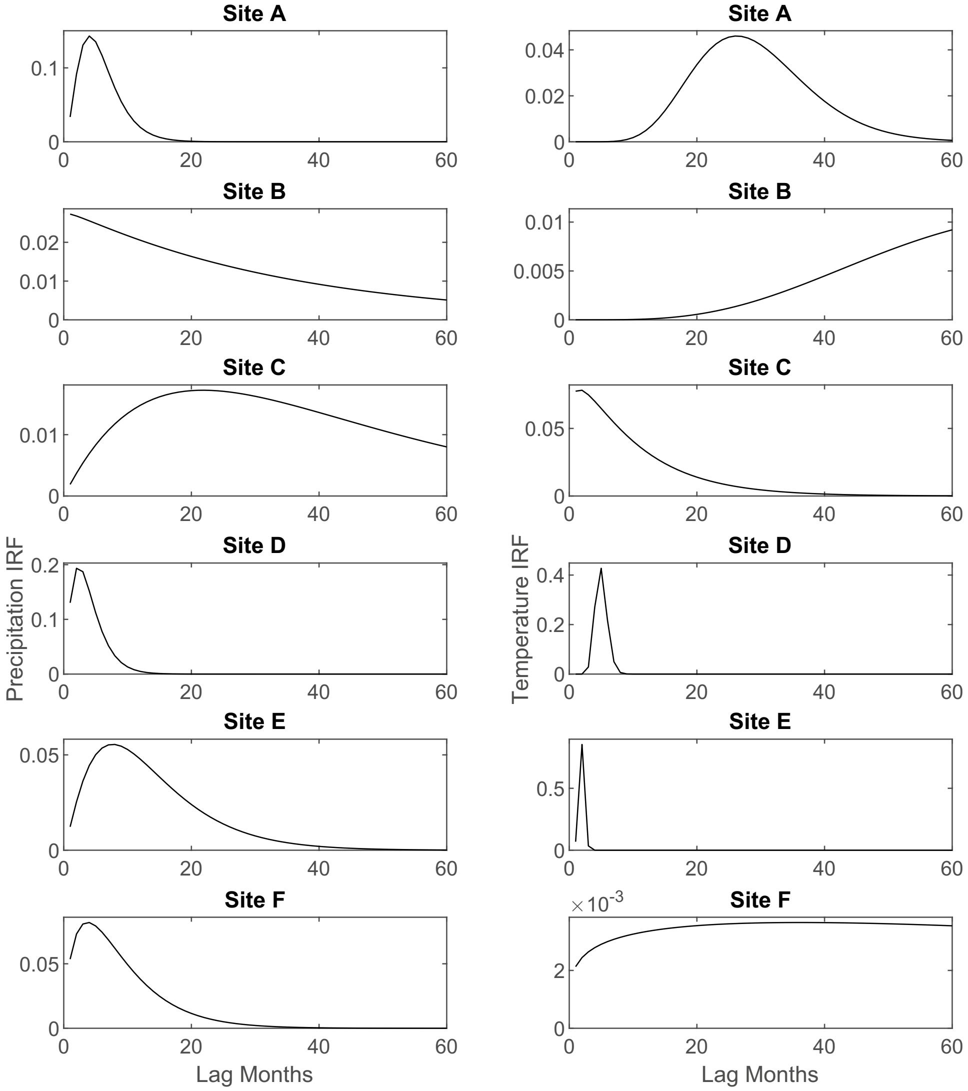

Fig. 3 Estimated GWL impulse response functions (IRFs) in response to precipitation (left panel) and temperature (right panel) at the six sites. Note that temperature IRFs for sites A, B, E and F do not include the Akaike information criterion (AIC) and are not included in the estimated linear mixed model

Validation of linear mixed models

The ten-fold cross-validation results (Table 3) indicate that the models for all six sites are approximately unbiased, with the largest ME occurring for site A and equal to 3% of the standard deviation of the observations for that site. All models with an exponential covariance function explain a substantial proportion of variation in the observations upon these cross-validation tests. The largest RMSE of 36% of the standard deviation of the observations occurs for site D. These relatively large errors reflect the rapid temporal changes in GWLs observed at this site. Similarly, all of the sites have a large concordance coefficient. For sites A, B and C, where the long-term trend explains a large proportion of the variation the concordance coefficients are greater than 0.98. The concordance coefficient at the other sites is only marginally smaller. Although site F has the largest random effects variance, the temporal correlation amongst these random effects can be used to accurately predict GWLs. The mean squared SPE upon 10-fold cross-validation are generally close to 1.0, indicating that the uncertainty of the predictions is reliably quantified. However, at site E, the mean squared SPE is a third larger than expected. This could indicate that the assumptions of the linear mixed model are not exactly honoured at this site—for instance, the exponential covariance function might not be sufficient to accurately approximate the auto-correlation of the random effects.

All models were re-estimated using a Matérn rather than exponential covariance function, and the cross-validation results are shown in Table 3 (note that all other figures and tables relate to models with an exponential covariance function). The more flexible covariance function did improve the mean SSPE for five of the six sites, including site E. There were also modest improvements in the RMSEs for these sites; however, the required computation times generally increased by a factor of more than five.

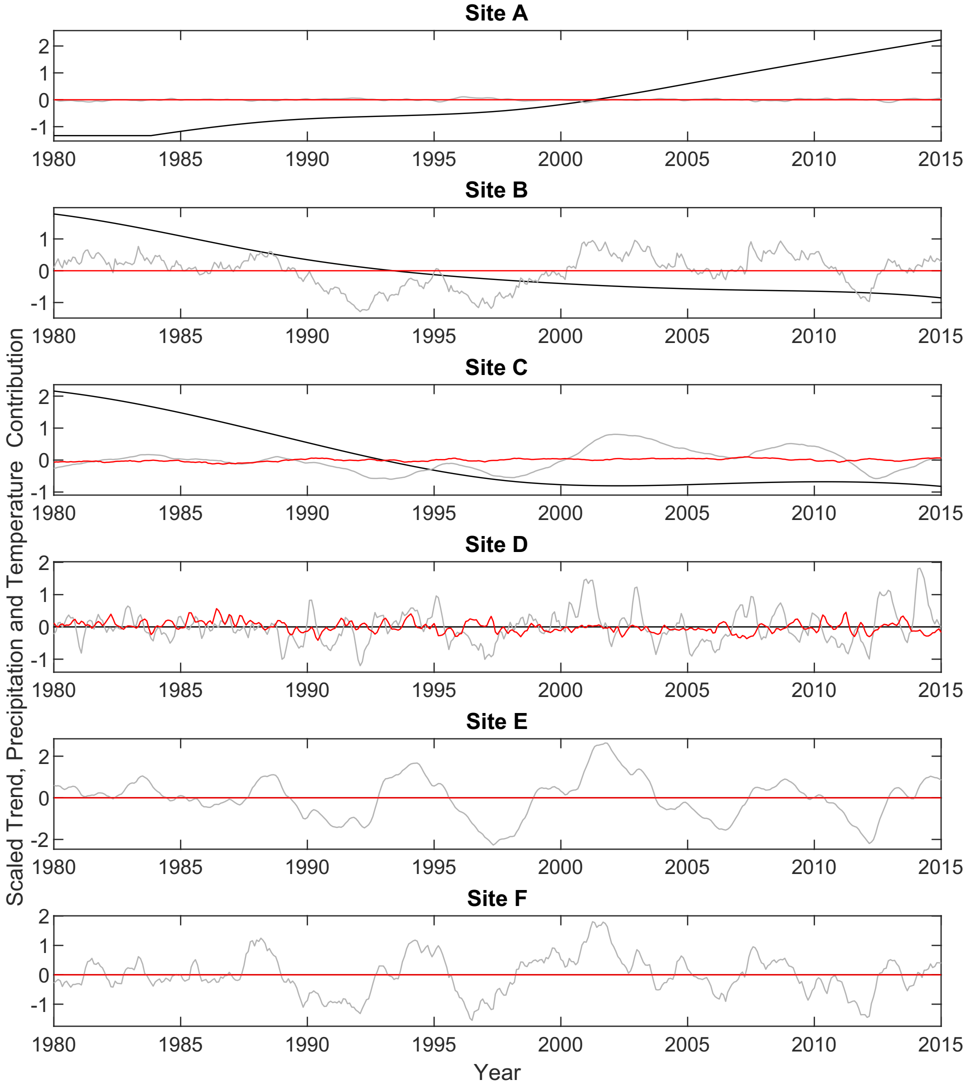

Fig. 4 Contributions from the long-term trend (black), precipitation (grey) and temperature (red) fixed effects to the scaled GWLs. Each contribution is shifted to zero mean

When performing 10-fold cross-validation upon this dataset, most predictions occur at times near to an observation that is available to the predictor. When the lags between the prediction and nearest observation times are controlled, there is some evidence of errors increasing with these lags (Fig. 5). This is most apparent for sites D, E and F where the random effects have a larger variance and the auto-correlation of the random effects has a greater influence on the predictions. However, large concordance coefficients (>0.75) are evident for all lags up to 15 months at all sites. Except for site E, the mean squared SPEs are close to 1.0 for all lags. The variation in these squared SPEs according to lag at site E is further evidence that a more general covariance function might be required at this site. When these validated GWLs are converted to SGI values, the RMSEs at sites A, B, and C are generally less than 0.5 for all lags (Fig. 4, bottom row). The SGI RMSEs are generally larger for sites D, E and F, but they are always less than 1, indicating that the infilling procedure is explaining some of the variations in SGI. The SGI RMSEs at sites D, E and F do increase with the time to nearest observation up to a maximum which occurs after a number of months of similar magnitude to the range of temporal auto-correlation for that site.

Prediction of GWLs and SGI

The predicted GWLs at each site for months where they were not observed are shown by the red dots in Fig. 2. From visual inspection these predictions appear to be consistent with the wider trends and patterns of variation within the time series. The flexibility of the linear mixed models is evident in that varying degrees of seasonality and responsiveness to precipitation dependent on the observed GWL variation at each site are included in the predictions. Where long periods of measurements are missing, any long-term trends in the data

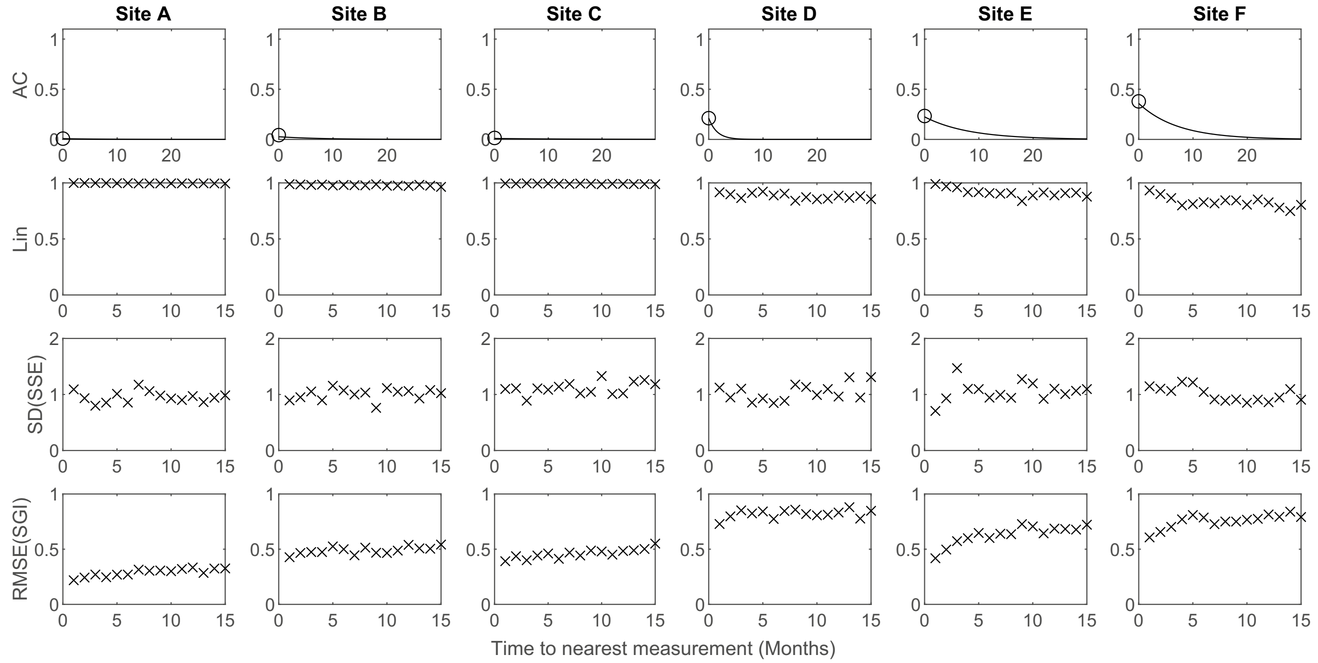

Fig. 5 The top row shows the estimated autocorrelation (AC) function of model residuals for each site. Circles indicate variance of residuals. The second row shows Lin's concordance coefficient between predicted and observed values upon validation plotted as a function of time to nearest observation; the third row shows mean standardised squared prediction errors (SSPE) upon validation plotted as a function of time to nearest observation; the bottom row shows root mean squared error of predicted standardised groundwater level index (SGI) plotted as a function of time to nearest observation

cannot be reliably predicted. This is most evident at site A between 1980 and 1984 when the long-term trend is assumed to remain constant although there are no data to support this assumption.

Where there are gaps in the GWL observations spanning multiple months, the estimated range of temporal auto-correlation for the site influences how quickly the predictions converge to the fixed effects terms—for example, the random effects for site B are only temporally correlated for short lags (Fig. 5) and the predictions shown in Fig. 2 tend the follow the fixed effects, particularly between 2002 and 2003. In contrast, the random effects for site F have a considerably longer temporal range and larger and longer lasting deviations between the fixed effects and the model predictions are apparent (e.g. in 1997).

Figure 6 shows the GWL time series when converted to SGI. Although it is not the aim of this paper to analyse or interpret the temporal changes in SGI between the six

Table 3 Autocorrelation (AC) model; mean error (ME); root mean square error (RMSE); Lin's concordance coefficient, mean standardised squared prediction errors (SSPE) upon 10-fold cross-validation of groundwater level model and computing time for each site. All sets of observations have been scaled to variance 1 and units are dimensionless

Borehole	AC model	ME	RMSE	Concordance	Mean SSPE	Computing time (secs)
A	Exponential	0.03	0.17	0.98	1.14	105
	Matérn	0.02	0.17	0.98	1.03	763
B	Exponential	0.00	0.15	0.99	1.10	128
	Matérn	–0.00	0.15	0.99	1.05	762
C	Exponential	0.00	0.08	1.00	1.03	130
	Matérn	–0.00	0.10	1.00	1.09	835
D	Exponential	0.01	0.36	0.93	0.92	98
	Matérn	0.01	0.35	0.93	1.01	389
E	Exponential	0.00	0.22	0.97	1.33	120
	Matérn	0.00	0.18	0.98	0.92	647
F	Exponential	0.02	0.30	0.95	1.13	115
	Matérn	0.02	0.27	0.96	1.08	680

sites, the figure demonstrates the utility of normalised and standardised groundwater level hydrographs. For example, it is evident from visual inspection that sites D, E and F show broadly similar temporal variations in SGI with low groundwater level stands (low SGI) in 1992–1993 rapidly transitioning to relatively high groundwater level stands (high SGI) in 1994. For large regional datasets with many SGI hydrographs, it is possible to undertake some form of clustering to identify sites with similar temporal variations in SGI and extract and characterise droughts from those collections of sites (Bloomfield et al. 2015; Marchant and Bloomfield 2018). Sites B and C show longer term-declines in SGI and site A an upward trend in SGI over the analysis period. Previously, Marchant and Bloomfield (2018) have inferred that such characteristics may indicate a degree of human influence on the SGI hydrographs (either long-term overexploitation or groundwater rebound) and screened such sites out of a large-sample analysis used for the characterisation of groundwater droughts. However, if long-term trends in groundwater status are the focus of future large-sample studies, then trend analysis of SGI time series, such as those shown in Fig. 6, would enable a quantitative assessment of changes in SGI between sites. The predicted GWLs and SGI values for the six sites reported here have been deposited in the NERC National Geoscience Data Centre (Bloomfield et al. 2022).

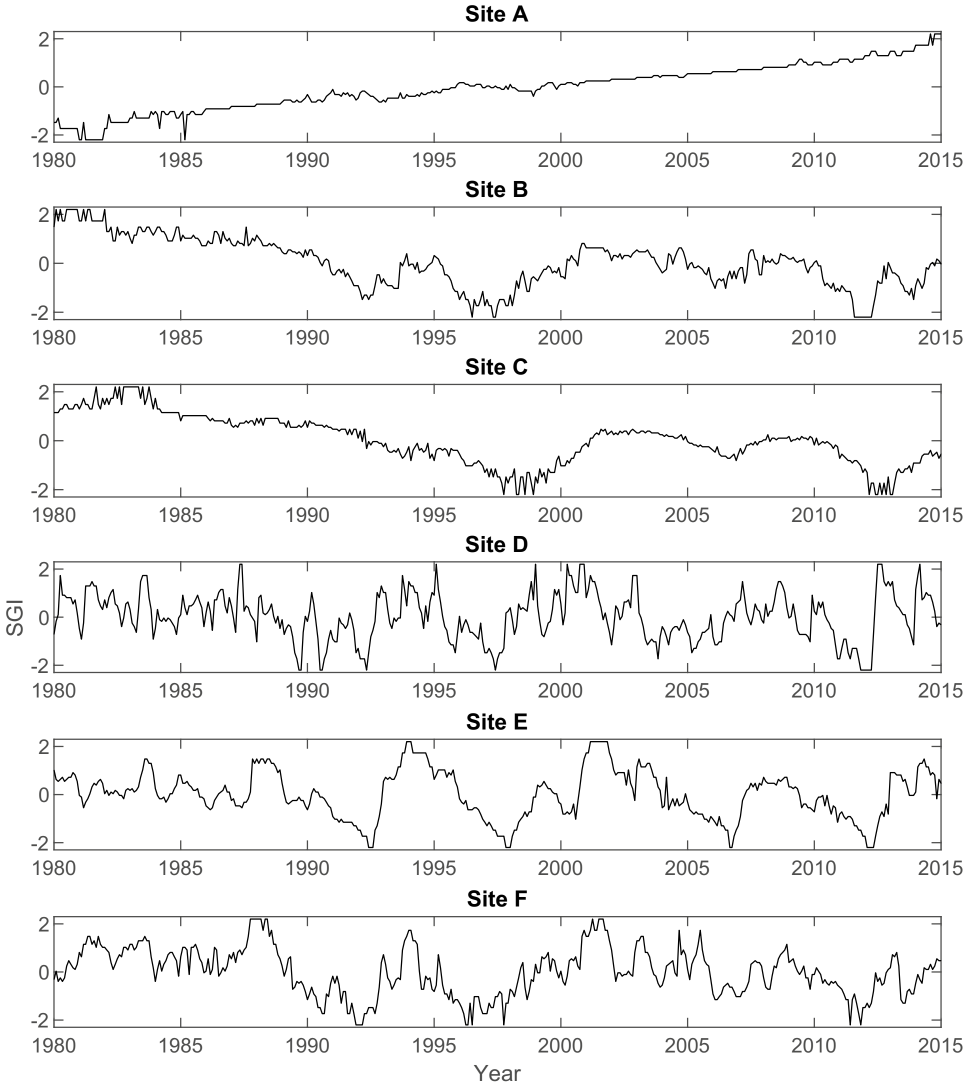

Fig. 6 Predicted standardised groundwater level index (SGI) time series for each of the six sites

Discussion

Effectiveness of linear mixed models in predicting GWLs

The linear mixed models presented here were computed in times less than or equal to 130 s and led to plausible predictions of GWLs on dates where they were not observed across all six sites. The models are flexible enough to represent disparate patterns of variation including long-term trends, seasonal behaviour of different magnitudes and responses to weather variables over different time scales. Little supervision of the modelling procedure is required and it appears to be suitable to be upscaled to much larger studies.

The validation tests indicated that all the models lead to large concordance coefficients between predicted and observed values and that the uncertainty of the predictions is quantified reasonably accurately. The largest discrepancy between realised errors and predicted uncertainty occurs at site E where squared SPEs are on average 33% higher than expected. This discrepancy can be reduced by using a more complex model with a Matérn rather than an exponential covariance function. For studies across a few sites, such modifications to the model would be desirable. However, the five-fold increase in computation times might be impractical for studies that include data from thousands of sites. The models continue to explain a substantial proportion of variation when the infilled GWLs are converted to SGI time series.

The reliability of the uncertainty quantification implies that the models can be used to identify observations that are inconsistent with the underlying variation in GWLs at each site. However, some caution should be applied when deciding how these outliers should be interpreted and treated. Some outliers are likely to be present because the linear mixed models are too simple to represent the full complexity of GWL variation. Therefore, a relatively large threshold on SPEs was adopted when identifying outliers to ensure that any removed values were very much inconsistent with the expected values and likely to be the result of errors or particularly transient processes.

The validation results appear to indicate that sufficient data are available at these six sites to calibrate a linear mixed model, infill the time series of observations, calculate the SGI series and understand the temporal variation in drought status of each site. The most obvious limitations in the models occurred when long-term trends were acting, but no observations were available at the start or end of the time series to constrain these trends. Therefore, the authors recommend that caution is applied in applying the approach if there are no observations within 12 months of the start or end of the study period. Validation errors do increase with the time lag between the nearest observation and the prediction time. Further recommendations include using only time series with no data gaps of more than 24 months. It is also necessary that more than 150 observations are available at the site so the geostatistical model can be reliably calibrated (Webster and Oliver 2007). However, these recommendations are necessarily subjective because the magnitude of errors at any site is a complex function of the proportion of variation explained by the fixed effects, the degree of temporal correlation of the random effects and the number and configuration of missing data. Once a model has been estimated, it should be validated using the approaches adopted in this paper to confirm that it is leading to informative predictions. The approaches described in this paper could be used for study periods of shorter duration than the 35 years considered here, provided the aforementioned recommendations are satisfied. However, for these shorter study periods, it is likely to be more challenging to identify relationships between GWLs and climatic variables, particularly if the impacts of these variables are integrated over many months.

Effectiveness of linear mixed models in interpreting drivers of GWL variation

The linear mixed models can potentially include four types of trends in the fixed effects, namely the seasonal variation, long-term trends, response to precipitation and response to temperature. The variable selection procedure indicates which of these have an impact at each site. However, when the proportion of variation explained by each term at each site is examined, some of the terms that pass a statistical test to be included in the model have a relatively small influence on GWLs. Therefore, the inclusion of a term in the model should not necessarily be seen as an indication of the importance of that term in controlling GWLs.

The approximate percentages of variance explained at each site indicate that the dominant terms vary between sites in line with the authors' expectations. There is substantial seasonal variation at sites D, E and F and to a lesser extent at site B. The long-term trend term dominates at sites A, B and C and precipitation explains more than 10% of the variation at all sites except site A. The temperature term is only included in the models from two sites and explains only a small proportion of variation. There is some weak evidence to suggest that the temperature term acts over a similar timescale to the precipitation term and perhaps corresponds to a time-varying proportion of precipitation that is lost to evapotranspiration. Data from regions with more extreme temperature variation are required to explore this proposal in more detail. It should be noted that the seasonal term, when it is included, will account for the effects of seasonal variation in temperature. The temperature term is therefore explaining variation from the seasonal norm.

It is possible to separately present the contribution of each fixed effect term in the model (Fig. 4). These plots indicated that the terms are not strongly correlated. The IRF plots (Fig. 3) provide a visualisation of how the hydrogeological system responds to precipitation and to a lesser extent temperature. Expert knowledge of the system and the natural or anthropogenic changes it has undergone are required to interpret the long-term trend components of variation. The modelling procedure can be performed with any combination of the different covariate terms described previously and, indeed, additional terms which may be introduced to the model in any order to reflect a conceptual model of the system being studied.

Some hydrogeological conditions are not currently included in the fixed effects model and, if these conditions occur at a particular site, their impacts will be included in the random effects. Therefore, the model cannot be used to readily interpret these impacts. Such conditions include tidally induced semidiurnal GWL fluctuations in coastal aquifer systems, nearby abstraction leading to short-term variation in GWLs and bodily connection with surface water bodies. If suitable sea level, abstraction and surface water level data were available then these sources of variation could be added to the fixed effects.

Detailed inspection of the precipitation IRFs can provide some insight into the processes controlling the relationship between precipitation and groundwater recharge (Calver 1997)—for example, peaks in the IFRs for small lags indicate rapid recharge perhaps caused by piston or by-pass flow (Al-Jaf et al. 2020).

Conclusions

The linear mixed modelling framework developed in this paper is suitable to infill GWL time series from thousands of boreholes as required in large-scale studies of groundwater drought. The approach is computationally efficient, flexible, and can identify unexpected and potentially erroneous observations. The models can accommodate seasonal variation, long-term trends and responses to precipitation and temperature over different temporal scales and examination of the relative contributions of each of these terms aids interpretation of the drivers of GWL variation. Validation of the model predictions confirms that the models explain a substantial proportion of GWL variation and that the uncertainty of these predictions is reliably quantified.

Acknowledgements This paper is published with the permission of the Executive Director of the British Geological Survey, part of UK Research and Innovation (UKRI).

Funding The work was part of the UK Natural Environment Research Council funded Groundwater Drought Initiative (award number NE/R004994/1).

Declarations

Conflicts of interest The authors have no conflicts of interest.

References

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F (eds) Proc. International Symposium on Information Theory, Ashkelon, Israel, June 1973, pp 267–281

Al-Jaf P, Smith M, Gunzel F (2020) Unsaturated zone flow processes and aquifer response time in the Chalk Aquifer, Brighton, South East England. Groundwater 59:381–395

Allen DJ, Brewerton LJ, Coleby LM, Gibbs BR, Lewis MA, MacDonald AM, Wagstaff SJ, Williams, AT (1997) The physical properties of major aquifers in England and Wales. British Geological Survey. BGS report no. WD/97/034, 333 pp. http://nora.nerc.ac.uk/id/eprint/13137/. Accessed July 2022

Ascott MJ, Mansour M, Bloomfield JP, Upton KA (2019) Analysis of the impact of hydraulic properties and climate change on estimations of borehole yields. J Hydrol 577:123998

Barker LJ, Hannaford J, Chiverton A, Svensson C (2016) From meteorological to hydrological drought using standardised indicators. Hydrol Earth Syst Sci 20:2483–2505

Beven K (2006) A manifesto for the equifinality thesis. J Hydrol 320:18–36

Birtles AR, Reeves MJ (1977) A simple effective method for the computer simulation of groundwater storage and its application in the design of water resource systems. J Hydrol 34:77–96

Bloomfield JP, Marchant BP (2013) Analysis of groundwater drought building on the standardised precipitation index approach. Hydrol Earth Syst Sci 17:4769–4787

Bloomfield JP, Allen DJ, Griffiths KJ (2009) Examining geological controls on baseflow index (BFI) using regression analysis: an illustration from the Thames Basin, UK. J Hydrol 373:164–176. https://doi.org/10.1016/j.hydrol.2009.04.025

Bloomfield JP, Marchant BP, Bricker SH, Morgan RB (2015) Regional analysis of groundwater droughts using hydrograph classification. Hydrol Earth Syst Sci 19(10):4327–4344

Bloomfield JP, Marchant BP, Brauns B (2022) Monthly groundwater level (GWL) and standardised groundwater levels for six sites in the UK illustrating the temporal interpolation of groundwater level hydrographs for regional drought analysis using mixed models. NERC EDS Natl Geosci Data Centre (Dataset). https://doi.org/10.5285/940c1b90-fef3-49c3-8be9-39b4b0078660

Brauns B, Cuba D, Bloomfield JP, Hannah DM, Jackson C, Marchant B, Heudorfer B, Van Loon AF, Bessière H, Thunholm B, Schubert G (2020) The Groundwater Drought Initiative (GDI): analysing and understanding groundwater drought across Europe. Proc Int Assoc Hydrol Sci 383:297–305

British Geological Survey (2020) WellMaster hydrogeological database. https://www.bgs.ac.uk/products/hydrogeology/wellmaster.html. Accessed Aug 2020

Calver A (1997) Recharge response functions. HESS 1:47–53

Dobson AJ (1990) An introduction to generalized linear models, 2nd edn. Boca Raton, FL

Folland CK, Hannaford J, Bloomfield JP, Kendon M, Svensson C, Marchant BP, Prior J, Wallace E (2015) Multi-annual droughts in the English Lowlands: a review of their characteristics and climate drivers in the winter half-year. Hydrol Earth Syst Sci 19(5):2353–2375

Gleeson T, Ingebritsen S (2016) Crustal permeability. Wiley, Chichester, UK

Gupta HV, Perrin C, Blöschl G, Montanari A, Kumar R, Clark M, Andréassian V (2014) Large-sample hydrology: a need to balance depth with breadth. Hydrol Earth Syst Sci 18:463–477

Harris I, Osborn TJ, Jones P et al (2020) Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. Sci Data 7:109. https://doi.org/10.1038/s41597-020-0453-3

Hellwig J, de Graaf IEM, Weiler M, Stahl K (2020) Large-scale assessment of delayed groundwater responses to drought. Water Resour Res 56:e2019WR025441

Heudorfer B, Haaf E, Stahl K, Barthel R (2019) Index-based characterization and quantification of groundwater dynamics. Water Resour Res 55:5575–5592

Jones HK, Morris BL, Cheney CS, Brewerton LJ, Merrin PD, Lewis MA, MacDonald AM, Coleby LM, Talbot JC, McKenzie AA, Bird MJ, Cunningham JE, Robinson V (2000) The physical properties of minor aquifers in England and Wales. Environment Agency R&D Publication 68, BGS report no. WD/00/004, British Geological Survey, 234 pp. http://nora.nerc.ac.uk/id/eprint/12663/. Accessed July 2022

Kazumba S, Oron G, Honjo Y, Kamiya K (2008) Lumped model for regional groundwater flow analysis. J Hydrol 359:131–140

Keating T (1982) A lumped parameter model of a chalk aquifer–stream system in Hampshire, United Kingdom. Ground Water 20:430–436

Kumar R, Musuuza JL, Van Loon AF, Teuling AJ, Barthel R, Broek JT, Mai J, Samaniego L, Attinger S (2016) Multiscale evaluation of the Standardized Precipitation Index as a groundwater drought indicator. Hydrol Earth Syst Sci 20:1117–1131

Lark RM, Cullis BR, Welham SJ (2006) On spatial prediction of soil properties in the presence of a spatial trend: the empirical best linear unbiased predictor (E-BLUP) with REML. Eur J Soil Sci 57:787–799

Li L, Wen Z, Wang Z (2016) Outlier detection and correction during the process of groundwater lever monitoring base on Pauta criterion with self-learning and smooth processing. In: Zhang L, Song X, Wu Y (eds) Theory, methodology, tools and applications for modeling and simulation of complex systems. Springer, Singapore, pp 497–503

Mackay JD, Jackson CR, Wang L (2014) A lumped conceptual model to simulate groundwater level time-series. Environ Model Softw 61:229–245

Marchant BP (2018) Model-based geostatistics. In: McBratney AB, Minasny B, Stockmann U (eds) Pedometrics: a system of quantitative soil information. Springer, Heidelberg, Germany

Marchant BP (2021) Using remote sensors to predict soil properties: radiometry and peat depth in Dartmoor, UK. Geoderma 403:115232

Marchant BP, Bloomfield JP (2018) Spatio-temporal modelling of the status of groundwater droughts. J Hydrol 564:397–413

McKee TB, Doesken NJ, Leist J (1993) The relationship of drought frequency and duration time scales. 8th Conference on Applied Climatology, Anaheim, CA, 17–22 January 1993, pp 179–184

Peterson TJ, Western AW (2014) Nonlinear time-series modelling of unconfined groundwater head. Water Resour Res 50:8330–8355

Peterson TJ, Western AW (2018) Statistical interpolation of groundwater hydrographs. Water Resour Res 54:4663–4680

Peterson TJ, Western AW, Cheng X (2018) The good, the bad and the outliers: automated detection of errors and outliers from groundwater hydrographs. Hydrogeol J 26:371–380

Schneider U, Becker A, Finger P, Meyer-Christoffer A, Ziese M (2018) GPCC full data monthly product version 2018 at 0.25°: monthly land-surface precipitation from rain-gauges built on GTS-based and historical data. https://doi.org/10.5676/DWD_GPCC/FD_M_V2018_025

Shabalala ZP, Moeletsi ME, Tongwane MI, Mazibuko SM (2019) Evaluation of infilling methods for time series of daily temperature data: case study of Limpopo Province, South Africa. Climate 7:86

Svensson C, Hannaford J, Prosdocimi I (2016) Statistical distributions for monthly aggregations of precipitation and streamflow in drought indicator applications. Water Resour Res 53:999–1018

Tremblay Y, Lemieux JM, Fortier R, Molson J, Therrien R, Therrien P, Comeau G, Poulin MC (2015) Semi-automated filtering of data outliers to improve spatial analysis of piezometric data. Hydrogeol J 23(5):851–868

Trichakis I, Burek P, de Roo A, Pistocchi A (2017) Towards a pan-European integrated groundwater and surface water model: development and applications. Environ Process 4(Suppl 1):S81–S93

Van Loon AF (2015) Hydrological drought explained. WIREs Water 2:359–392

Vicente-Serrano SM, Beguería S, López-Moreno JI (2010) A multiscalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index. J Clim 23(7):1696–1718

von Asmuth JR, Bierkens MFP (2005) Modeling irregularly spaced residual series as a continuous stochastic process. Water Resour Res 41:W12404

Von Asmuth JR, Bierkens MFP, Maas K (2002) Transfer function-noise modelling in continuous time using predefined impulse response functions. Water Resour Res 38(12):23-1–23-12

Webster R, Oliver MA (2007) Geostatistics for environmental scientists, 2nd edn. Wiley, Chichester, UK

Wood SN (2017) Generalized additive models: an introduction with R, 2nd edn. CRC, Boca Raton, FL

Zaghiyan MR, Eslamian S, Gohari A, et al (2021) Temporal correction of irregular observed intervals of groundwater level series using interpolation techniques. Theor Appl Climatol. https://doi.org/10.1007/s00704-021-03666-1

Zhang Y, Thorburn PJ (2022) Handling missing data in near real-time environmental monitoring: a system and a review of selected methods. Futur Gener Comput Syst 128:63–72

Zhou Y, Li W (2011) A review of regional groundwater flow modelling. Geosci Front 2:205–214