ACCEPTED MANUSCRIPT

Seasonal forecasting of groundwater levels in principal aquifers of the United Kingdom

J.D. Mackay[a]*, C.R. Jackson[a], A. Brookshaw[b], A.A. Scaife[b], J. Cook[c] and R.S. Ward[a].

[a] British Geological Survey, Environmental Science Centre, Keyworth, Nottingham NG12 5GG, UK

[b] Met Office, FitzRoy Road, Exeter, EX1 3PB, UK

[c] ISA Lille, Lille, France.

* Corresponding author: Tel.: +44 1159 363100; fax: +44 1159 363200; email: joncka@bgs.ac.uk

Abstract

To date, the majority of hydrological forecasting studies have focussed on using medium-range (3 to 15 days) weather forecasts to drive hydrological models and make predictions of future river flows. With recent developments in seasonal (1 to 3 months) weather forecast skill, such as those from the latest version of the UK Met Office global seasonal forecast system (GloSea5), there is now an opportunity to use similar methodologies to forecast groundwater levels in more slowly responding aquifers on seasonal timescales. This study uses seasonal rainfall forecasts and a lumped groundwater model to simulate groundwater levels at 21 locations in the United Kingdom up to three months into the future. The results indicate that the forecasts have skill; outperforming a persistence forecast and demonstrating reliability, resolution and discrimination. However, there is currently little to

ACCEPTED MANUSCRIPT

gain from using seasonal rainfall forecasts over using site climatology for this type of application. Furthermore, the forecasts are not able to capture extreme groundwater levels, primarily because of inadequacies in the driving rainfall forecasts. The findings also show that the origin of forecast skill, be it from the meteorological input, groundwater model or initial condition, is site specific and related to the groundwater response characteristics to rainfall and antecedent hydro-meteorological conditions.

Keywords

Seasonal forecasting; ensemble forecasting; groundwater level forecasting; AquiMod; GloSea5.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

1. Introduction

Often a cleaner and more reliable source of drinking water than surface reservoirs, groundwater aquifers comprise the world's largest freshwater resource and provide resilience to climate extremes which may increase in frequency with future climate change (Alley et al., 2002; Mishra and Singh, 2010; Sukhija, 2008). Under prolonged dry climatic conditions groundwater drought can develop, often characterised by significantly low groundwater levels which persist for months to years (Lanen and Peters, 2000; Marsh et al., 2007). This may lead to the drying up of significant water-bearing wells and the degradation of ecologically important rivers and springs. Conversely, lasting wet conditions can induce anomalously high groundwater levels resulting in persistent flooding, potentially at large economic cost (Huntingford et al., 2014; Pinault et al., 2005; Upton and Jackson, 2011). Proper management of these resources is vital to ensure their sustainability and to reduce the risk and impacts from groundwater level extremes.

One possible way forward is to forecast future groundwater levels so that management strategies can be employed in advance of likely future events. However, these approaches generally require some insight into future weather patterns and an understanding of site-specific hydrogeological characteristics that control the non-linear groundwater discharge response to changes in groundwater levels (Eltahir and Yeh, 1999; Moore and Bell, 1999). This paper attempts to do this by using state-of-the-art seasonal weather forecasts to drive a series of groundwater models to forecast groundwater levels up to three months into the future.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

The majority of groundwater level forecasting studies have been conducted using black-box modelling approaches (Jakeman et al., 2006) whereby an empirical relationship between groundwater level time-series and one or more predictor variables is found using an optimization algorithm (Sahu, 2003). Typically, meteorological covariates, including rainfall and temperature, are used because these perturb groundwater recharge fluxes. Flow through the unsaturated zone and saturated aquifer can slow the response of groundwater level to rainfall events (Alley et al., 2002). Accordingly, a suitable characterisation of this lagged response may be sufficient for forecasting future groundwater levels in aquifers, given up-to-date weather data.

The most widely used method to characterise the lagged response of groundwater levels to meteorological predictor variables is the non-parametric Artificial Neural Network (ANN), a flexible tool that is able to implement multiple statistical models to replicate patterns in time-series (Maier and Dandy, 2000). Daliakopoulos et al. (2005) used neural networks to forecast monthly groundwater levels in a highly heterogeneous alluvial aquifer in Crete, Greece. Trichakis et al. (2009) also used ANNs to forecast the change in hydraulic head in a complex karstic limestone aquifer in Greece which proved to be accurate up to a 90-day lead time. Taormina et al. (2012) forecast groundwater levels on an hourly time-step for a flashy shallow coastal aquifer in the Venice lagoon and found that they could accurately reproduce groundwater depths for several months ahead. These, along with other studies that have used ANNs (Nourani et al., 2008; Sreekanth et al., 2009; Ying et al., 2014) all show significant forecasting skill months into the future. However, there are two key limitations with these approaches: i) not all aquifers exhibit a significant lagged response to antecedent weather; and ii) to forecast more than one time-step ahead these studies used retrospective observed meteorological predictor variables which would not be available ahead of time.

Tsanis et al. (2008) recognised the second issue and adapted the work of Daliakopoulos et al. (2005) to include a precipitation projection model which, if used in combination with seasonally averaged temperature data, could simulate groundwater levels up to 30 months ahead, achieving a $R^2 > 0.9$. It should be noted, however, that it is likely that this high correlation score largely reflects the model's ability to capture a downward groundwater level trend induced by steady abstractions in the dry season. Even so, it does demonstrate the possibility of using meteorological forecasts to extend the lead time of real-time groundwater level projections.

Alternative black box methods such as support vector machines (Behzad et al., 2010; Suryanarayana et al., 2014; Vapnik, 1999; Yoon et al., 2011) and wavelet decompositions (Adamowski and Chan, 2011; Maheswaran and Khosa, 2013; Partal and Kişi, 2007) have also been used for groundwater level forecasting in the past with promising levels of skill. Mendicino et al. (2008) took a different approach by using a simple conceptual distributed water balance model to derive average groundwater storage over the most southern peninsular of Italy, the outputs of which were used to derive a groundwater drought index. They found that due to the persistence of low groundwater levels in the summer months, droughts could be forecast months prior to their occurrence based on model simulations of the current groundwater storage.

ACCEPTED MANUSCRIPT

While these studies have shown some skill, the relative infancy of groundwater level forecasting science becomes apparent when compared to the abundance of studies focussed on forecasting other hydrological variables such as river discharge for flood forecasting (see Cloke and Pappenberger (2009) and Cuo et al. (2011) for two comprehensive reviews of these applications). Here, forecasters are not afforded the luxury of long response times to prior weather patterns. At the catchment scale, river flow response time to rainfall is typically of the order of minutes to hours. As such, forecasters drive their hydrological models with medium-range weather forecast products from numerical weather prediction (NWP) centres, which typically offer lead times of 3 to 15 days. These extended lead times may allow water resource managers and contingency planners to implement mitigation strategies in advance of extreme events. Of course, the benefit of increased lead time comes at a cost; namely that these meteorological products are inherently uncertain due to the non-linear, chaotic nature of the atmosphere (Lorenz, 1963). In response, river flow forecasters now adopt probabilistic methodologies that incorporate this uncertainty rather than relying on a single deterministic forecast. A popular approach that couples probability with determinism is ensemble forecasting (Lewis, 2005) whereby a number of deterministic weather forecasts with differing initial conditions are used to drive the hydrological model. If these realisations are assumed independent and of the same random process, it is possible to assign probabilities to the occurrence or exceedance of given flow thresholds. This probabilistic, ensemble-based approach provides more consistent and skilful outlooks from which users can manage risks more effectively (Addor et al., 2011; Buizza, 2008). One may also cascade other uncertainties, such as those associated with the hydrological model parameterisation, through the forecasting system (Beven, 2006; Pappenberger et al., 2005; Zappa et al., 2010; Zappa et al., 2011). A well

ACCEPTED MANUSCRIPT

established approach for this is the Generalised Likelihood Uncertainty Estimation (GLUE) method (Beven and Binley, 1992; Beven and Binley, 2013), whereby an informal likelihood function is used to weight an ensemble of behavioural models. It should be noted, however, that due to the computational burden, such approaches for real-time hydrological forecasting applications are still not widely used today.

The response of groundwater levels to rainfall generally operate on longer time scales (days to months) than river flows. As such, strategies to mitigate an imposing groundwater drought, for example, can only be properly formulated with a good understanding of the likely future groundwater levels over a similar time scale. Here, longer-range weather forecasts on the scale of months would be required, like those produced by the latest version of the Met Office global seasonal forecast system (GloSea5) which are now showing increased skill up to a three month lead time (Scaife et al., 2014). To date, however, the majority of seasonal forecasting studies have been undertaken by the river flow forecasting community. Yossef et al. (2012) investigated the potential for forecasting monthly and seasonal river flow extremes in 20 large river basins around the world by driving the global hydrological model, PCR-GLOBWB (Sperna Weiland et al., 2010) with observed meteorological forcing data. They found that they could capture observed flood and drought events given skilful meteorological inputs. More recently, Svensson et al. (2015) used GloSea5 seasonal rainfall forecasts to drive a 1 km resolution water balance model (Bell et al., 2013) and forecast winter (December-January-February) river flows across the UK. The forecasts correlated with observed winter river flows with a median correlation score of 0.45. They also found a clear geographical contrast in the source of predictability whereby the initial condition was the strongest source of predictability in the more

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

permeable, baseflow-dominated catchments of south-east England, and the skill was much more dependent on the meteorological forcing data for the flashy catchments in the north-west of Great Britain. The role of river flow response characteristics on seasonal forecast skill was also found to be important for global seasonal river flow forecasting by Yossef et al. (2013). Indeed, contrasting response characteristics to rainfall can also be found in groundwater level time-series (e.g. see the work of Bloomfield and Marchant, 2013), and these are likely to influence the sensitivity of groundwater level forecasts to the meteorological forcing data.

To summarise, skilful forecasts of groundwater levels would provide useful information to water resource managers and contingency planners which could help to mitigate hazards such as groundwater flooding and drought, both of which can lead to social, economic and environmental degradation. Experience gained from the river flow forecasting community shows that skilful ensemble hydrological forecasts can be generated using driving data from medium-range NWP models. However, because aquifers generally respond to prevailing weather patterns over a number of months, the insight gained over a 15-day lead time may be small. This has led most studies to rely on the lagged response of groundwater levels to past weather patterns to make forecasts. However, it may be possible to extend the skilful forecast lead time using seasonal weather forecast products to drive groundwater models, an approach that is already showing some skill in river flow forecasting experiments.

This paper presents a new probabilistic groundwater level forecasting approach that utilises state-of-the-art GloSea5 multi-member seasonal forecasts of rainfall produced by the UK Met Office to drive a series of groundwater models up to three months into the future. A

ACCEPTED MANUSCRIPT

parsimonious lumped conceptual groundwater model, AquiMod (Mackay et al., 2014), which simulates groundwater levels at observation boreholes has been used. The models have been calibrated to simulate groundwater level time-series at 21 locations across the UK and in different aquifers with contrasting hydrogeological properties and response characteristics to rainfall. The skill of the groundwater level forecasts is evaluated over the four UK seasons using a 14-year sequence of GloSea5 rainfall reforecasts. For comparison, reforecasts using rainfall climatology and observed rainfall have also been evaluated. Consideration of the catchment response characteristics and their influence on forecast skill are also made. From these analyses, this study seeks to provide a first evaluation of the potential for national, real-time seasonal groundwater level forecasting.

2. Methodology

2.1. Study catchments

In total, 21 groundwater catchments, each with an observation borehole and associated groundwater level record were selected for this study from a database of 181 groundwater level time-series held in the National Groundwater Level Archive (Marsh and Hannaford, 2008). They were selected because: i) they are situated in unconfined aquifers for which the AquiMod groundwater model is best suited; ii) they are located away from any significant groundwater abstractions; and iii) they have continuous monthly groundwater level records that cover the operational 14-year GloSea5 reforecast period from March 1996 to February 2010 (MacLachlan et al., 2014) and at least 15 years of data prior to this for model calibration. The boreholes penetrate into some of the UK's principal aquifers including the Cretaceous Chalk and Lower Greensand, the Jurassic and Magnesian Limestone and the

ACCEPTED MANUSCRIPT

Permo-Triassic Sandstone (Figure 1). Between 16 and 34 years of continuous groundwater level data were available for model calibration.

Figure 2 shows the raw groundwater level time-series for four of the observation boreholes. Also included are the groundwater level auto-correlation plots and the rainfall-groundwater level cross-correlation plots. It can be seen that groundwater level fluctuations contrast between the catchments. For example, Ashton Farm shows a sinusoidal pattern with relatively uniform amplitude while the New Red Lion borehole shows more variable amplitude with multiple winter peaks. The West Dean cross-correlation plot shows the highest correlation between groundwater and rainfall at a lag of zero, indicating a very rapid and flashy response. This is in contrast to the smooth Heathlanes hydrograph which exhibits relatively small seasonal variability, but more pronounced inter-annual fluctuations. The auto-correlation and cross-correlation plots for this site indicate significant persistence in levels and a much longer response time to rainfall. Also note that because this borehole is in a high storage Sandstone aquifer, annual groundwater levels typically fluctuate by only 0.5 m. In contrast, the water table at New Red Lion in the low porosity Jurassic Limestone aquifer can vary by as much as 20 m in one year.

2.2. AquiMod

AquiMod takes monthly rainfall and potential evapotranspiration (PET) driving data and uses conceptual hydrological equations to simulate the downward movement of water through the soil and unsaturated zone and the lateral flow and subsequent discharge of groundwater through the saturated zone (Figure 3). A soil module divides rainfall between evapotranspiration, runoff and soil drainage. The soil drainage is attenuated through the

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

unsaturated zone using a Weibull distribution transfer function, before reaching the saturated zone as groundwater recharge. Discharge from the saturated zone is calculated using a Darcy flux equation. The reader is referred to Mackay et al. (2014) for a more comprehensive description of the underlying theory and model code.

The AquiMod code was chosen for this study because it was designed specifically for simulating groundwater levels at observation boreholes. It includes in built Monte Carlo parameter sampling, has a small computational burden and also allows the user to incorporate different saturated zone model structures with variable levels of complexity. Mackay et al. (2014) showed that this model can efficiently capture the non-linear groundwater level dynamics in a range of hydrogeological settings. They also showed that a two or three layer aquifer representation is generally most efficient and these structures have been adopted in this study (Figure 3).

2.3. Model calibration

The models were driven with observed monthly rainfall and PET data and calibrated against observed groundwater levels prior to the reforecast period. Rainfall data were obtained from the national 5 km gridded dataset held by the UK Met Office National Climate Information Centre (Perry et al., 2009). This is comprised of rain gauge data interpolated onto a regular grid using inverse-distance weighting. PET data were extracted from the Met Office Rainfall and Evapotranspiration Calculation System (MORECS) dataset (Field, 1983) which uses synoptic station data in conjunction with a modified version of the Penman-Monteith equation to determine the monthly average PET rate on a 40 km grid of over the UK (Monteith and Unsworth, 2008).

ACCEPTED MANUSCRIPT

Following the methodology of Mackay et al. (2014), eight of the possible 16 model parameters were fixed based on known catchment characteristics while the remaining were used as calibration parameters (Table 1). A Monte Carlo procedure was used to randomly select 10^6 unique parameter sets from a user-defined parameter space for each model structure. Here we considered the uncertainty in model structure and parameter selection by adopting the GLUE methodology and using the well established Nash-Sutcliffe efficiency (NSE) score (Bennett et al., 2013; Nash and Sutcliffe, 1970) as the informal likelihood measure. Only those models that exceeded a NSE score of 0.5 were deemed behavioural. Those that did not achieve this were assigned a likelihood of zero.

Between 1780 and 2470 behavioural models were obtained for the 21 study catchments achieving a maximum efficiency (NSE_{max}) between 0.71 and 0.94 and a containment ratio (CR) (Xiong and O'Connor, 2008), which specifies the percentage of observations captured within specified upper and lower prediction bounds, between 65.3 and 97.8% when using the GLUE 95% confidence interval (Table 2). Figure 4 shows the observed and simulated groundwater levels with the GLUE 95% prediction bounds for Bussels, where AquiMod achieved the highest NSE_{max}, and Therfield Rectory, where AquiMod scored the lowest NSE_{max}. It can be seen that the GLUE prediction bounds for Bussels contain almost all of the observations and the best model closely replicates the timing and seasonality in the hydrograph including the pronounced 1976 drought period. For Therfield Rectory, AquiMod captures the timing and seasonality of the hydrograph and most of the observations during the drought of 1973. However, it fails to capture the rapid recession in 1992, and some of

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

the peak levels observed in 1961, 1979 and 1988. These deficiencies are considered in more detail in the discussion.

2.4. Reforecast climate data

Monthly rainfall inputs for the 14-year reforecast period were taken from the GloSea5 model. These comprised four reforecasts per year representing the four seasons: i) spring March-April-May (MAM); ii) summer June-July-August (JJA); iii) autumn September-October-November (SON); iv) winter December-January-February (DJF). Each consisted of an ensemble of one, two and three month ahead rainfall, averaged over the UK. The GloSea5 winter and summer forecasts were made up of 24 ensemble members while the spring and autumn forecasts were made up of 12 ensemble members. All were downscaled to the catchment scale using linear models defined by ordinary least squares regression between observed catchment rainfall and observed UK average rainfall. Figure 5a and Figure 5b show the relationship between seasonal UK average and seasonal catchment rainfall for the Ashton Farm and New Red Lion observation boreholes. The fitted linear regression models are shown by the solid black lines. It can be seen that for Ashton Farm, a linear approximation of the scale relationship is satisfactory, giving an R^2 score of 0.51, while for New Red Lion, this approximation is less adequate, where the model only explains 31% of the variance. In general, however, the linear regression models demonstrated a good fit, with a mean R^2 score of 0.46 across the study catchments. These models were then used to downscale the GloSea5 forecasts of UK average rainfall for each catchment. The downscaled GloSea5 seasonal rainfall forecasts for Ashton Farm showed the most skill, where the ensemble mean seasonal rainfall correlated with the observed catchment rainfall with an R^2 of 0.44 (Figure 5c). In contrast, the downscaled GloSea5 seasonal rainfall forecasts for New

ACCEPTED MANUSCRIPT

Red Lion showed negligible correlation with the observed catchment rainfall (Figure 5d). Overall, the skill of the downscaled rainfall forecasts was low with a mean R^2 of 0.19 across the study catchments.

For each seasonal reforecast at a given location, the population of behavioural models were run for two years using observed rainfall and PET data to initialise the soil and unsaturated zone in the models. Their initial groundwater levels were fixed to the latest observation. The models were then run for a further three months using the rainfall and PET data described above, producing an ensemble of predictions with n*m members, where n is the number of behavioural models, and m is the rainfall ensemble size. The predicted groundwater level probability density function was then constructed using the predefined GLUE likelihoods and assuming equal probability of occurrence for each rainfall ensemble member.

2.5. Skill analysis

When evaluating forecast skill, it is often useful to establish categorical events for which the observed and forecast frequencies can be compared. Here, three categorical events were chosen for each catchment; below, near and above normal groundwater levels, defined by monthly terciles from the observed groundwater level data. Jolliffe and Stephenson (2012) detail a vast number of forecast verification metrics. We have chosen to use four quantitative metrics which assess different aspects of forecast skill for a given categorical event including:

ACCEPTED MANUSCRIPT

1. **Frequency bias:** The ratio of the total number of forecast occurrences to the total number of observed events. Here, the forecast event was defined as that which had the highest forecast probability.
2. **Reliability:** The consistency between the forecast probabilities and the observed relative frequencies. Here, a negatively oriented reliability score derived from the decomposition of the brier score (Murphy, 1973) has been used.
3. **Relative operating characteristic (ROC) score:** This measures the capacity to correctly discriminate between the occurrence and non-occurrence of an event. A value greater than 0.5 indicates that the hit rate exceeds the false alarm rate.
4. **Continuous ranked probability score (CRPS):** Calculated as the integrated square difference between the cumulative distributions of the forecasts and observations. This is a probabilistic generalisation of the mean absolute error.

We chose to convert the CRPS into a skill score, the CRPSS, by comparing the groundwater level forecasts to a reference persistence forecast. A persistence-type benchmark was deemed the most rigorous test given that hydrogeological memory can serve as a potential source of skill. We evaluated three different benchmarks against historical observed groundwater levels including i) persisting the latest observed groundwater level; ii) perturbing the latest observed groundwater level using the monthly mean changes in groundwater levels taken from historical data; and iii) persisting the percentile location of the initial groundwater level in the distribution of historical groundwater levels for that month over the following three months (i.e. if the initial condition was near-normal, the forecast for the subsequent three months would remain in this category). We found that

ACCEPTED MANUSCRIPT

the third approach was the best, consistently outperforming the other two benchmarks and so this was deemed the most rigorous test for forecast skill.

To complement the benchmark tests, the groundwater models have also been driven with two other meteorological inputs including: i) an unskilful rainfall ensemble made up of re-sampled observed catchment data; and ii) a best case deterministic rainfall input using observed data.

3. Results

It is known that groundwater levels respond to rainfall differently between the catchments. It is therefore likely that the models will also respond differently. This is examined in the first part of the results by undertaking a sensitivity analysis of the models. The results from this are used to organise the models into a number of response type groups. Note here, and in the text that follows, the term model refers to the population of behavioural models for a given catchment rather than a single model realisation. The remainder of this section analyses the skill of the forecasts for each of the response groups, first by using the skill metrics outlined above and then by analysing a selection of forecast time-series plots.

3.1. Groundwater level response to rainfall

It is postulated that because of the contrasting response characterises to rainfall across the catchments, the calibrated models will exhibit different sensitivities to rainfall over the three month forecast horizon. Understanding these sensitivities is important because they influence the added value of using seasonal rainfall forecasts to simulate future groundwater levels.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

A relative measure of sensitivity to rainfall has been derived for each of the calibrated models. To do this, each model was spun-up using observed rainfall and PET and then run for three months using six arbitrary synthetic rainfall inputs ranging from 0 to 5 mm d^{-1}. This process was repeated using each of the months in the reforecast sequence as the initial condition. The sensitivity was then calculated for each month as the range of the six groundwater level forecasts, normalised with respect to the model specific yield. This normalisation step accounts for the different storage properties of each model to allow for easier inter-model comparison.

Figure 6a shows how the model sensitivity to rainfall changes over the reforecast period for one, two and three month simulations for the Rockley observation borehole in the Chalk aquifer. As would be expected, the sensitivity increases with lead time as the influence of the initial condition diminishes, but there is also a seasonal cycle with peak sensitivity during the winter and considerably reduced sensitivity in the summer months. Given that the climate data for the forecasts are fixed, these variations are a result of perturbations in the initial conditions only. This can be explained by the initial soil moisture deficit (SMD) conditions in the model (Figure 6b) which generally develop in the warmer summer months and must be satisfied before recharge (Figure 6d) is initiated. In the winter months the SMD is small and so small changes in the rainfall input can significantly perturb the modelled groundwater level. Despite this, the sensitivity can increase as the soil moisture deficit develops (for example see year 2003 boxed in Figure 6). This behaviour can be attributed to the initial groundwater level condition, which shows to be receding, and the quadratic groundwater discharge response to a unit rise in groundwater head. In other words, as the

ACCEPTED MANUSCRIPT

groundwater level recedes, the discharge response to an influx of recharge is smaller, and so the sensitivity increases.

Similar seasonal fluctuations in sensitivity were observed for all of the study catchments, but the magnitude varied substantially. The reason for this is likely to be multifaceted, but it can be attributed primarily to the different model response times to rainfall. It is possible to investigate this by considering the calibrated unsaturated zone Weibull distribution transfer function in AquiMod which spreads the flux of water from the soil zone to the water table over a number of months. This transfer function can be evaluated at lags covering the three month forecast horizon to define a model response characteristic, P, which specifies the percentage of modelled effective rainfall that reaches the water table over this period. Figure 7a shows that this value ranges between 20 – 95% and that the relationship between P and the derived model sensitivities can be approximated with an exponential curve (R^2 = 0.79) that shows that as P increases, the model sensitivity to rainfall also increases. The permeability of each catchment is also likely to influence the sensitivity to rainfall. Indeed, a closer fit is obtained if the model sensitivity is normalised by the catchment baseflow index (BFI) (Figure 7b), taken from Marsh and Hannaford (2008), which defines the proportion of effective rainfall that contributes to groundwater flow. Furthermore, the calibrated model sensitivities also correlate well with an independent inference of the response time to rainfall for each catchment estimated by the peak lead lag correlation (CC_{max}) between observed rainfall and de-seasonalised groundwater levels (Figure 7c), obtaining an R^2 of 0.76.

ACCEPTED MANUSCRIPT

These findings demonstrate that the more pronounced the AquiMod lagging mechanism, the less sensitive the three month simulations are to the choice of rainfall input. A similar relationship between the sensitivity and an independent estimation of the peak groundwater level response time to rainfall for each borehole, further indicates that the catchment response time has a clear influence on the sensitivity, and therefore is also likely to influence the skill of the forecasts. Consequently, the catchments have been split into three equally sized groups representing slowly responding ($3 \leq CC_{max} \leq 10$), moderately responding ($1 \leq CC_{max} \leq 2$) and quickly responding ($0 \leq CC_{max} \leq 1$) groundwater catchments. These are indicated in Figure 7a-c by the circles, squares and triangles respectively and the analyses in the subsequent sections are conducted using this grouping.

3.2. Skill metrics

For the purpose of this skill analysis the reforecasts have been subdivided into 36 different assessment groups for which an independent assessment of skill has been conducted. These groups comprise the three categorical events, the four seasons and the three groundwater response groups. Figure 8 shows the four skill measurements for all of the assessment groups using the three different climate inputs.

The frequency bias ranges between 0.61 and 0.5 (Figure 8a). In the summer (JJA), there is a consistent under forecasting of below normal levels which is mainly offset by a positive frequency bias for near normal events. The winter (DJF) forecasts show the opposite pattern, under forecasting above normal events and over forecasting below normal events. The fact that groundwater levels tend to peak in the winter and trough in the summer indicates that there is a tendency for the forecasts to miss the groundwater level extremes.

ACCEPTED MANUSCRIPT

Indeed, on average, the above and below normal events show negative frequency biases of -0.04 and -0.11 respectively while the near normal event category shows a positive frequency bias of 0.13. This deficiency cannot be attributed to the driving rainfall data as all assessment groups demonstrate that they are insensitive to this, except for the quickly responding catchments in winter and autumn (SON) where using the best case observed climate reduces the bias by approximately half. This insensitivity is also apparent in the other skill metrics, indicating that the skill or lack of it stems more from the model than the rainfall input in most situations.

Generally, the forecasts are more reliable when predicting above and below normal events than near normal events, especially during winter and autumn and during the summer for the quickly responding catchments (Figure 8b). Figure 9 shows the reliability diagrams for the quickly responding catchments in winter. It can be seen that for the above and below normal events the reliability curves follow the line of perfect reliability closely indicating good consistency between the forecast probabilities and observed relative frequencies. In contrast, there is a tendency for the forecasts to predict closer to base rate probabilities (0.33) for the near normal events as indicated by the flat reliability curves which imply a lack of forecast resolution. This is reflected in the ROC scores (Figure 8c) which are smaller on average for the near normal events indicating that the forecasts are less efficient at discriminating these events. Even so, all of the ROC scores obtained were greater than 0.5 showing that the number of hits exceeded the number of false alarms. The forecasts were also able to discriminate below normal events with an average ROC score of 0.87 using the downscaled GloSea climate which is particularly encouraging.

ACCEPTED MANUSCRIPT

The ROC scores also demonstrate a clear relationship with the catchment response times where the less sensitive, slowly responding catchments have greater discrimination capacity than the quickly responding catchments. However, this again appears to be an artefact of the model skill rather than to do with the sensitivity to the rainfall input. Even so, the use of observed climate consistently improves the discrimination capacity of the forecasts, particularly for the quickly responding catchments where improvements of up to 0.14 are shown.

From the 36 assessment groups, 35 return a positive CRPSS when using the climatology rainfall input (Figure 8d). This indicates that even climatology yields forecasts that are a better predictor of groundwater levels than a persistence forecast. Slightly fewer (31) of the groups return a positive CRPSS using the observed rainfall and only 30 when using the downscaled GloSea data. All suggest that the forecasts consistently outperform the persistence approach.

3.3. Time-series analysis

Finally, the forecasts have been evaluated over three time periods which contain important historical events including: i) the onset and persistence of below normal levels in 1996 and 1997, a period where many parts of the UK experienced groundwater drought; ii) the subsequent transition back to normal levels in 1997 and 1998, broadly associated with the end of the drought; and (iii) the onset and peak of above normal levels in the winter of 2000/2001, a period where many boreholes recorded their highest ever levels and where there was widespread groundwater flooding, particularly in the Chalk of south-east England. Figure 10 shows the number of catchments in each response category that successfully

ACCEPTED MANUSCRIPT

forecast each event using the three different climate inputs. A forecast was only deemed a success if all of the observed groundwater levels were contained within the forecast uncertainty bounds as defined by the limits of the ensemble. In addition, Figure 11 displays time-series plots for several of the study catchments over these events which have been used to compare the observed groundwater levels (black dots) against the ensemble mean forecasts (thick dashed lines) using the GloSea5 and observed rainfall inputs. The uncertainty bounds (thin dashed lines) are also shown.

The forecasts were least effective at capturing the high levels of winter 2000/2001 when using the downscaled GloSea and climatology rainfall, but showed significant improvements when driven with observed data. This is demonstrated in Figure 11a where the observed initial groundwater rise (time steps one to three) at the quickly responding New Red Lion borehole, is only replicated by the ensemble mean forecast when using the observed rainfall input. The forecast using the downscaled GloSea rainfall does not capture this due to underestimating the seasonal rainfall by almost 130 mm. Note that the GloSea forecast is able to capture the peak groundwater level at the fourth time step. This is partly because this corresponds to the one month ahead forecast, and therefore the model was initialised at the above normal levels from the previous time step.

For the below normal levels of 1996 and 1997, the choice of climate has less impact on the success rate which is demonstrated by the New Red Lion reforecast in Figure 11b. It can be seen here that regardless of the rainfall input, the ensemble mean overestimates the groundwater levels and the uncertainty bounds do not capture the gradual recession of the hydrograph and even extend into the above normal range between time steps five and six.

ACCEPTED MANUSCRIPT

This insensitivity could be explained by the large soil moisture deficit that would likely develop over this period. Therefore, the forecasts are much more reliant on the skill of the model, which in this case does not capture the groundwater discharge and subsequent hydrograph recession adequately.

Some catchments, such as the quickly responding Bussels catchment (Figure 11c) did demonstrate significant sensitivity to the rainfall input during this drought period. Here, it can be seen that when using the observed rainfall, the ensemble mean follows the persistent low groundwater levels closely, but when using the downscaled GloSea rainfall, the ensemble mean forecast actually predicts a sharp rise in groundwater level almost back to normal conditions (time steps seven to nine) due to the downscaled GloSea forecast overestimating rainfall by 100 mm for this period.

The highest overall success rates using the downscaled GloSea inputs were recorded for the return to normal levels in 1997 and 1998. For the moderately responding Rockley observation borehole (Figure 11d), it can be seen that the two ensemble mean forecasts using the GloSea and observed rainfall inputs are similar. Furthermore, both capture all of the observations in their uncertainty bounds which was observed for most of the catchments for this period.

4. Discussion

This study has demonstrated that skilful seasonal forecasts of groundwater levels at observation boreholes can be generated by using seasonal weather forecasts to drive parsimonious conceptual groundwater models. The forecasts were proficient at

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

discriminating between below, near and above normal future groundwater levels and they consistently outperformed a reference persistence forecast system. They also demonstrated good reliability, particularly for the seasonal forecasts of spring groundwater levels. These positive attributes have also been demonstrated for the quickly responding catchments, indicating that the skill can extend beyond the peak response time of these groundwater systems.

The skill of the forecasts originates from a combination of the driving climate data, the groundwater models and the initial groundwater level condition. For those catchments where groundwater levels respond more slowly to rainfall, the groundwater models and the initial conditions have a stronger influence on the forecast skill than the rainfall input. However, there is no clear indication that the sensitivity to the rainfall input directly affects the forecast skill. Rather, the relationship between groundwater level response time to rainfall and forecast skill appears to be primarily controlled by the groundwater model efficiency. Indeed, when conducting this work, we could find no apparent correlation between skill and geographical location like, for example, the work of Svensson et al. (2015). However, we suggest that with a larger sample size of boreholes, and by evaluating the forecast skill at longer lead times, where meteorological driving data plays a more crucial role in the forecast skill, such relationships may become more apparent. Indeed, while all of the response groups demonstrated forecast skill, it remains to be seen at what lead time this skill diminishes.

The origin of forecast skill also changes as a response to antecedent hydro-meteorological conditions. Here, it was found that when a large soil moisture deficit is developed during the

ACCEPTED MANUSCRIPT

model spin up and initialisation, the subsequent forecasts are less sensitive to the rainfall input. As such, we noted that for the summer forecasts, the skill derives mainly from the groundwater models and their internal hydrogeological memory. This has potential implications because some of the models have shown deficiencies, such as poor representation of the hydrograph recession, which materialised as forecast errors. Some of these deficiencies are likely to result from imperfect model calibration, errors in the meteorological input data and observed groundwater levels, or from inadequacies of the model structure and parameters. In this study, no account of input error was made, but we did acknowledge some of the model uncertainties by using an equifinality of acceptable model structures and parameter sets. Of course, this approach in itself may also propagate forecasting errors. For example, the choice of the NSE as a measure of model likelihood was subjective, and as with any objective function, is subject to undesirable properties that are likely to manifest themselves as modelling errors (Smith et al., 2008). There is also evidence that model appropriateness depends strongly on hydro-climatic conditions (Herman et al., 2013) and that it may be beneficial to develop better suited limits of acceptability which can be relaxed dynamically as a mean to implicitly account for input errors (Liu et al., 2009).

In contrast to the forecasts issued following dry conditions, during winter, when the soil is generally more saturated, the forecasts are more sensitive to the driving rainfall data, and as such the meteorological forecasts play a more crucial role in the skill of the groundwater level forecasts. The winter forecasts using the downscaled GloSea and climatology rainfall inputs both consistently outperformed the persistence approach, although it should be noted that using the downscaled GloSea5 rainfall data showed no significant improvement over using the site climatology inputs, and in some cases showed to be a worse rainfall

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

predictor. This is perhaps not surprising given that UK rainfall has complex spatio-temporal signatures that make deriving robust downscaling transformations difficult. Certainly, the linear downscaling model employed showed to be inadequate for some sites, and improving this should be a high priority for improving site-specific hydrological forecasts like these. However, it may be possible to improve this using more sophisticated non-linear downscaling and post processing techniques which have shown to be effective for medium-range ensemble streamflow forecasts (Verkade et al., 2013). Further data assimilation could also provide enhancements in skill through dynamic updating of state variables and forecast errors, although to date there is limited evidence that this is useful for seasonal or groundwater level forecasting applications (Liu et al., 2012). There are also other seasonal weather forecasting models which could be used for these types of applications, such as the System 4 from the European Centre for Medium-range Weather Forecasts (ECMWF) which has shown “marginally useful” degrees of reliability over Northern Europe (Weisheimer and Palmer, 2014).

It is important to note that the interpretation of skill in this study is primarily based on analysis of the verification metrics and by comparing the forecasts to the benchmark results. Pappenberger et al. (2015) compared a range of benchmarks for medium-range river flow forecasting and they note that the best benchmarks are the ones that are hardest to beat. While considerable effort was made to select appropriate benchmarks and avoid reporting “naïve” skill, it should be noted that the persistence benchmark used is less skilful for those boreholes that exhibit significant inter-annual groundwater level fluctuations, and so there is likely to be positive bias in the CRPSS reported for the more slowly responding catchments. Certainly, an equivalent thorough examination of benchmark performance to

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

that of Pappenberger et al. (2015) is also needed for seasonal groundwater level forecasting.

It is also important to consider that, the verification metrics used in this study only give an average indication of the forecast's ability to reliably discriminate between the occurrence of below, near and above normal levels over the 14-year reforecast sequence. When looking at the extreme 2000/2001 high groundwater level event specifically, only two of the 21 groundwater level forecasts were able to capture it within their uncertainty bounds when using the downscaled GloSea and climatology inputs. For the 1996/1997 drought period, the timing of the return to normal conditions could only be predicted when using observed rainfall data. This is an important issue, as it is arguably extreme events like these that, if foreseeable, would provide the most economic, environmental and societal benefit. That of course is not to say that these forecasts are not useful; on the contrary the Environment Agency in England, for example, routinely use measures of aquifer levels relative to normal conditions to inform agricultural communities about future prospects for spray irrigation and this approach can be used to help aid decision making processes for these needs. It does however mean that if we wish to forecast the initiation or end of extreme events on a seasonal time scale at the catchment or borehole resolution, then further enhancements in the skill and the use of seasonal rainfall forecasts are required.

5. Conclusions

Using seasonal weather forecasts to drive 21 conceptual groundwater models, this study has shown that skilful seasonal forecasts of groundwater levels at observation boreholes can be generated up to three months into the future. Site-specific groundwater level

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

response characteristics to rainfall result in contrasting sensitivities to the driving rainfall input across the study catchments. These sensitivities have also shown to be strongly controlled by prevailing weather conditions, where dry conditions tend to result in forecasts that are strongly controlled by the groundwater model, and wet conditions result in forecasts that are much more reliant on good driving rainfall data. This has important implications for where the skill or lack of it derives from, and more importantly, where future improvements can be made. There are clearly issues with correctly forecasting extreme groundwater levels which are primarily due to lack of skill in the driving rainfall data. In particular it is recommended that future work should focus these aspects:

1. Investigate the best practice for data assimilation, downscaling and post processing of seasonal weather forecasts for hydrological forecast applications.
2. Compare the use of different seasonal forecast products such as those produced by the ECMWF System 4 model.
3. Examine the maximum skilful forecast lead time for different aquifers in relation to their response characteristics to rainfall.

6. Acknowledgements

Mackay, Jackson and Ward were supported by core science funds of the Natural Environment Research Council (NERC) British Geological Survey's (BGS) Groundwater Science and Environmental Modelling Directorates. Scaife was supported by the Joint DECC/Defra Met Office Hadley Centre Climate Programme (GA01101) and Brookshaw was supported by the UK Public Weather Service research program. The groundwater level data for this study were taken from the NERC BGS National Groundwater Level Archive. The climate data were made available by the NERC Centre for Ecology and Hydrology, and the

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

UK Met-Office. Mackay, Jackson and Ward publish with the permission of the Executive Director of the British Geological Survey.

7. References

Adamowski, J. & Chan, H. F. 2011. A wavelet neural network conjunction model for groundwater level forecasting. *Journal of Hydrology*, 407**,** 28-40.

Addor, N., Jaun, S., Fundel, F. & Zappa, M. 2011. An operational hydrological ensemble prediction system for the city of Zurich (Switzerland): skill, case studies and scenarios. *Hydrol. Earth Syst. Sci.*, 15**,** 2327-2347.

Alley, W. M., Healy, R. W., LaBaugh, J. W. & Reilly, T. E. 2002. Flow and Storage in Groundwater Systems. *Science*, 296**,** 1985-1990.

Behzad, M., Asghari, K. & Coppola, E. 2010. Comparative Study of SVMs and ANNs in Aquifer Water Level Prediction. *Journal of Computing in Civil Engineering*, 24**,** 408-413.

Bell, V. A., Davies, H. N., Kay, A. L., Marsh, T. J., Brookshaw, A. & Jenkins, A. 2013. Developing a large-scale water-balance approach to seasonal forecasting: application to the 2012 drought in Britain. *Hydrological Processes*, 27**,** 3003-3012.

Bennett, N. D., Croke, B. F. W., Guariso, G., Guillaume, J. H. A., Hamilton, S. H., Jakeman, A. J., Marsili-Libelli, S., Newham, L. T. H., Norton, J. P., Perrin, C., Pierce, S. A., Robson, B., Seppelt, R., Voinov, A. A., Fath, B. D. & Andreassian, V. 2013. Characterising performance of environmental models. *Environmental Modelling & Software*, 40**,** 1-20.

Beven, K. 2006. A manifesto for the equifinality thesis. *Journal of Hydrology*, 320**,** 18-36.

Beven, K. & Binley, A. 1992. The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes*, 6**,** 279-298.

Beven, K. & Binley, A. 2013. GLUE: 20 years on. *Hydrological Processes***,** n/a-n/a.

Bloomfield, J. P. & Marchant, B. P. 2013. Analysis of groundwater drought building on the standardised precipitation index approach. *Hydrol. Earth Syst. Sci.*, 17**,** 4769-4787.

Boorman, D. B., Hollis, J. M. & Lilly, A. 1995. Report No. 126 Hydrology of soil types: a hydrologically-based classification of the soils of the United Kingdom. Institute of Hydrology, Wallingford, UK.

ACCEPTED MANUSCRIPT

Buizza, R. 2008. The value of probabilistic prediction. *Atmospheric Science Letters*, 9**,** 36-42.

Cloke, H. L. & Pappenberger, F. 2009. Ensemble flood forecasting: A review. *Journal of Hydrology*, 375**,** 613-626.

Cuo, L., Pagano, T. C. & Wang, Q. J. 2011. A Review of Quantitative Precipitation Forecasts and Their Use in Short- to Medium-Range Streamflow Forecasting. *Journal of Hydrometeorology*, 12**,** 713-728.

Daliakopoulos, I. N., Coulibaly, P. & Tsanis, I. K. 2005. Groundwater level forecasting using artificial neural networks. *Journal of Hydrology*, 309**,** 229-240.

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

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

Herman, J. D., Reed, P. M. & Wagener, T. 2013. Time-varying sensitivity analysis clarifies the effects of watershed model formulation on model behavior. *Water Resources Research*, 49**,** 1400-1414.

Huntingford, C., Marsh, T., Scaife, A. A., Kendon, E. J., Hannaford, J., Kay, A. L., Lockwood, M., Prudhomme, C., Reynard, N. S., Parry, S., Lowe, J. A., Screen, J. A., Ward, H. C., Roberts, M., Stott, P. A., Bell, V. A., Bailey, M., Jenkins, A., Legg, T., Otto, F. E. L., Massey, N., Schaller, N., Slingo, J. & Allen, M. R. 2014. Potential influences on the United Kingdom's floods of winter 2013/14. *Nature Clim. Change*, 4**,** 769-777.

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

Jolliffe, I. T. & Stephenson, D. B. 2012. Forecast Verification: A Practitioner's Guide in Atmospheric Science. Wiley, Chichester, UK.

Lanen, H. A. J. & Peters, E. 2000. Definition, Effects and Assessment of Groundwater Droughts. *In:* VOGT, J. & SOMMA, F. (eds.) *Drought and Drought Mitigation in Europe.* Springer Netherlands.

Lewis, J. M. 2005. Roots of Ensemble Forecasting. *Monthly Weather Review*, 133**,** 1865-1885.

Liu, Y., Freer, J., Beven, K. & Matgen, P. 2009. Towards a limits of acceptability approach to the calibration of hydrological models: Extending observation error. *Journal of Hydrology*, 367**,** 93-103.

Liu, Y., Weerts, A. H., Clark, M., Hendricks Franssen, H. J., Kumar, S., Moradkhani, H., Seo, D. J., Schwanenberg, D., Smith, P., van Dijk, A. I. J. M., van Velzen, N., He, M., Lee, H., Noh, S. J., Rakovec, O. & Restrepo, P.

ACCEPTED MANUSCRIPT

2012. Advancing data assimilation in operational hydrologic forecasting: progresses, challenges, and emerging opportunities. *Hydrol. Earth Syst. Sci.*, 16**,** 3863-3887.
Lorenz, E. N. 1963. Deterministic Nonperiodic Flow. *Journal of the Atmospheric Sciences*, 20**,** 130-141.
Mackay, J. D., Jackson, C. R. & Wang, L. 2014. A lumped conceptual model to simulate groundwater level time-series. *Environmental Modelling & Software*, 61**,** 229-245.
MacLachlan, C., Arribas, A., Peterson, K. A., Maidens, A., Fereday, D., Scaife, A. A., Gordon, M., Vellinga, M., Williams, A., Comer, R. E., Camp, J., Xavier, P. & Madec, G. 2014. Global Seasonal forecast system version 5 (GloSea5): a high-resolution seasonal forecast system. *Quarterly Journal of the Royal Meteorological Society***,** n/a-n/a.
Maheswaran, R. & Khosa, R. 2013. Long term forecasting of groundwater levels with evidence of non-stationary and nonlinear characteristics. *Computers & Geosciences*, 52**,** 422-436.
Maier, H. R. & Dandy, G. C. 2000. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environmental Modelling & Software*, 15**,** 101-124.
Marsh, T., Cole, G. & Wilby, R. 2007. Major droughts in England and Wales, 1800–2006. *Weather*, 62**,** 87-93.
Marsh, T. J. & Hannaford, J. 2008. UK Hydrometric Register. Centre for Ecology and Hydrology, Wallingford, UK.
Mendicino, G., Senatore, A. & Versace, P. 2008. A Groundwater Resource Index (GRI) for drought monitoring and forecasting in a mediterranean climate. *Journal of Hydrology*, 357**,** 282-302.
Mishra, A. K. & Singh, V. P. 2010. A review of drought concepts. *Journal of Hydrology*, 391**,** 202-216.
Monteith, J. L. & Unsworth, M. H. 2008. Principles of Environmental Physics: Third Edition. Elsevier, London, UK.
Moore, R. J. & Bell, V. A. 1999. Incorporation of groundwater losses and well level data in rainfall-runoff models illustrated using the PDM. *Hydrol. Earth Syst. Sci.*, 6**,** 25-38.
Murphy, A. H. 1973. A New Vector Partition of the Probability Score. *Journal of Applied Meteorology*, 12**,** 595-600.
Nash, J. E. & Sutcliffe, J. V. 1970. River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, 10**,** 282-290.

ACCEPTED MANUSCRIPT

Nourani, V., Mogaddam, A. A. & Nadiri, A. O. 2008. An ANN-based model for spatiotemporal groundwater level forecasting. *Hydrological Processes*, 22**,** 5054-5066.

Pappenberger, F., Beven, K. J., Hunter, N. M., Bates, P. D., Gouweleeuw, B. T., Thielen, J. & de Roo, A. P. J. 2005. Cascading model uncertainty from medium range weather forecasts (10 days) through a rainfall-runoff model to flood inundation predictions within the European Flood Forecasting System (EFFS). *Hydrol. Earth Syst. Sci.*, 9**,** 381-393.

Pappenberger, F., Ramos, M. H., Cloke, H. L., Wetterhall, F., Alfieri, L., Bogner, K., Mueller, A. & Salamon, P. 2015. How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction. *Journal of Hydrology*, 522**,** 697-713.

Partal, T. & Kişi, Ö. 2007. Wavelet and neuro-fuzzy conjunction model for precipitation forecasting. *Journal of Hydrology*, 342**,** 199-212.

Perry, M., Hollis, D. & Elms, M. 2009. The Generation of Daily Gridded Datasets of Temperature and Rainfall for the UK. Exeter, UK.

Pinault, J. L., Amraoui, N. & Golaz, C. 2005. Groundwater-induced flooding in macropore-dominated hydrological system in the context of climate changes. *Water Resources Research*, 41**,** W05001.

Sahu, B. K. 2003. Time Series Modelling in Earth Sciences. A.A. Balkema, Netherlands.

Scaife, A. A., Arribas, A., Blockley, E., Brookshaw, A., Clark, R. T., Dunstone, N., Eade, R., Fereday, D., Folland, C. K., Gordon, M., Hermanson, L., Knight, J. R., Lea, D. J., MacLachlan, C., Maidens, A., Martin, M., Peterson, A. K., Smith, D., Vellinga, M., Wallace, E., Waters, J. & Williams, A. 2014. Skillful long-range prediction of European and North American winters. *Geophysical Research Letters*, 41**,** 2014GL059637.

Smith, P., Beven, K. J. & Tawn, J. A. 2008. Informal likelihood measures in model assessment: Theoretic development and investigation. *Advances in Water Resources*, 31**,** 1087-1100.

Sperna Weiland, F. C., van Beek, L. P. H., Kwadijk, J. C. J. & Bierkens, M. F. P. 2010. The ability of a GCM-forced hydrological model to reproduce global discharge variability. *Hydrol. Earth Syst. Sci.*, 14**,** 1595-1621.

Sreekanth, P. D., Geethanjali, N., Sreedevi, P. D., Ahmend, S., Ravi Kumar, N. & Kamala Jayanthi, P. D. 2009. Forecasting groundwater level using artificial neural networks *Current Science*, 96**,** 933-939.

Sukhija, B. S. 2008. Adaptation to climate change: strategies for sustaining groundwater resources during droughts. *Geological Society, London, Special Publications*, 288**,** 169-181.

ACCEPTED MANUSCRIPT

Suryanarayana, C., Sudheer, C., Mahammood, V. & Panigrahi, B. K. 2014. An integrated wavelet-support vector machine for groundwater level prediction in Visakhapatnam, India. *Neurocomputing*, 145, 324-335.

Svensson, C., Brookshaw, A., Scaife, A. A., Bell, V. A., Mackay, J. D., Jackson, C. R., Hannaford, J., Davies, H. N., Arribas, A. & Stanley, S. 2015. Long-range forecasts of UK winter hydrology. *Environmental Research Letters*, 10, 064006.

Taormina, R., Chau, K.-w. & Sethi, R. 2012. Artificial neural network simulation of hourly groundwater levels in a coastal aquifer system of the Venice lagoon. *Engineering Applications of Artificial Intelligence*, 25, 1670-1676.

Trichakis, I. C., Nikolos, I. K. & Karatzas, G. P. 2009. Optimal selection of artificial neural network parameters for the prediction of a karstic aquifer's response. *Hydrological Processes*, 23, 2956-2969.

Tsanis, I. K., Coulibaly, P. & Daliakopoulos, I. N. 2008. Improving groundwater level forecasting with a feedforward neural network and linearly regressed projected precipitation. *Journal of Hydroinformatics*, 10, 317-330.

Upton, K. A. & Jackson, C. R. 2011. Simulation of the spatio-temporal extent of groundwater flooding using statistical methods of hydrograph classification and lumped parameter models. *Hydrological Processes*, 25, 1949-1963.

Vapnik, V. N. 1999. An overview of statistical learning theory. *Neural Networks, IEEE Transactions on*, 10, 988-999.

Verkade, J. S., Brown, J. D., Reggiani, P. & Weerts, A. H. 2013. Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales. *Journal of Hydrology*, 501, 73-91.

Weisheimer, A. & Palmer, T. N. 2014. On the reliability of seasonal climate forecasts.

Xiong, L. & O'Connor, K. M. 2008. An empirical method to improve the prediction limits of the GLUE methodology in rainfall–runoff modeling. *Journal of Hydrology*, 349, 115-124.

Ying, Z., Wenxi, L., Haibo, C. & Jiannan, L. 2014. Comparison of three forecasting models for groundwater levels: a case study in the semiarid area of west Jilin Province, China. *Journal of Water Supply*, 63, 671-683.

Yoon, H., Jun, S.-C., Hyun, Y., Bae, G.-O. & Lee, K.-K. 2011. A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer. *Journal of Hydrology*, 396, 128-138.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Yossef, N. C., van Beek, L. P. H., Kwadijk, J. C. J. & Bierkens, M. F. P. 2012. Assessment of the potential forecasting skill of a global hydrological model in reproducing the occurrence of monthly flow extremes. *Hydrol. Earth Syst. Sci.*, 16**,** 4233-4246.

Yossef, N. C., Winsemius, H., Weerts, A., van Beek, R. & Bierkens, M. F. P. 2013. Skill of a global seasonal streamflow forecasting system, relative roles of initial conditions and meteorological forcing. *Water Resources Research*, 49**,** 4687-4699.

Zappa, M., Beven, K. J., Bruen, M., Cofiño, A. S., Kok, K., Martin, E., Nurmi, P., Orfila, B., Roulin, E., Schröter, K., Seed, A., Szturc, J., Vehviläinen, B., Germann, U. & Rossa, A. 2010. Propagation of uncertainty from observing systems and NWP into hydrological models: COST-731 Working Group 2. *Atmospheric Science Letters*, 11**,** 83-91.

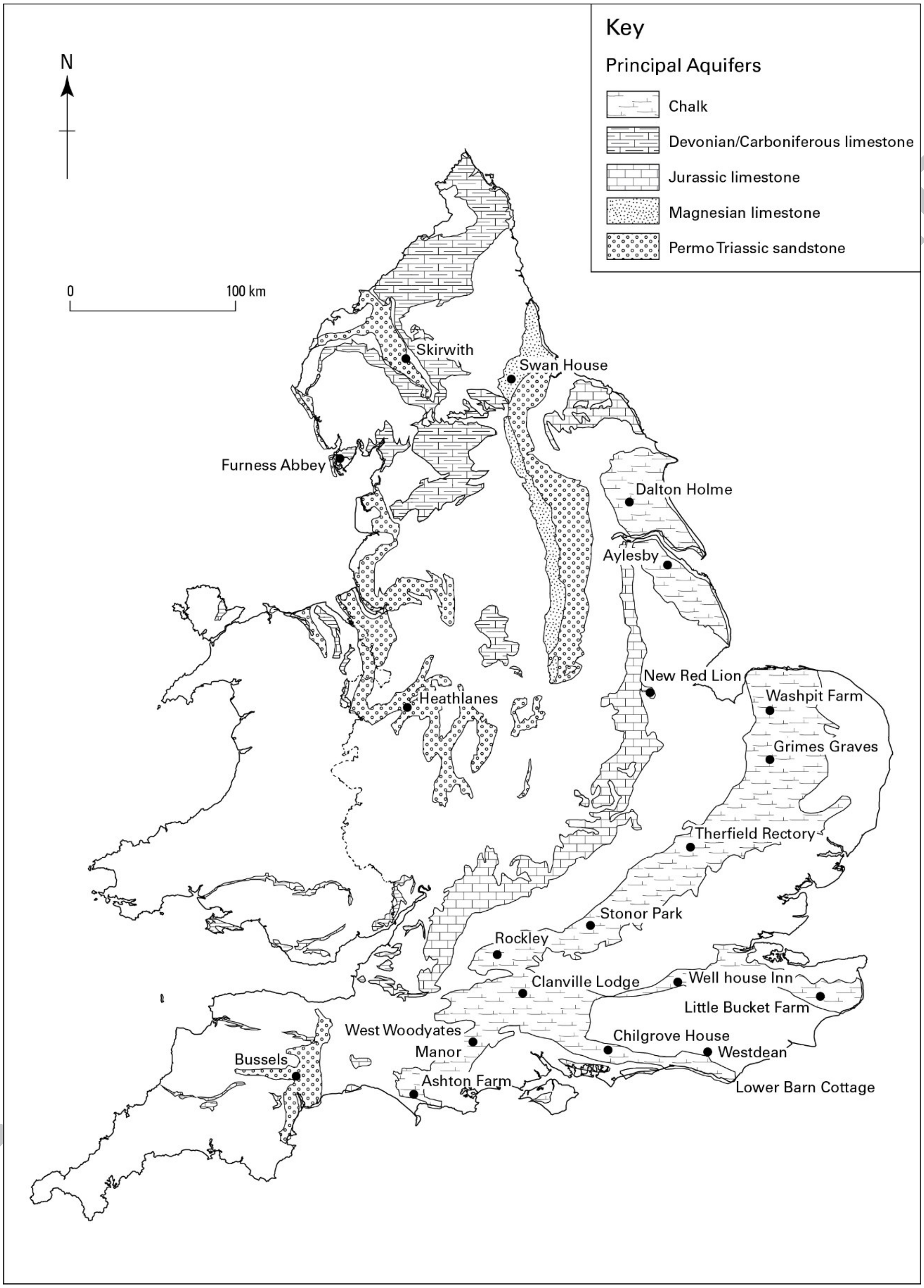
Zappa, M., Jaun, S., Germann, U., Walser, A. & Fundel, F. 2011. Superposition of three sources of uncertainties in operational flood forecasting chains. *Atmospheric Research*, 100**,** 246-262.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

893 **1. Figures**



894

895 *Figure 1: Observation borehole locations across the principal aquifers of the UK.*

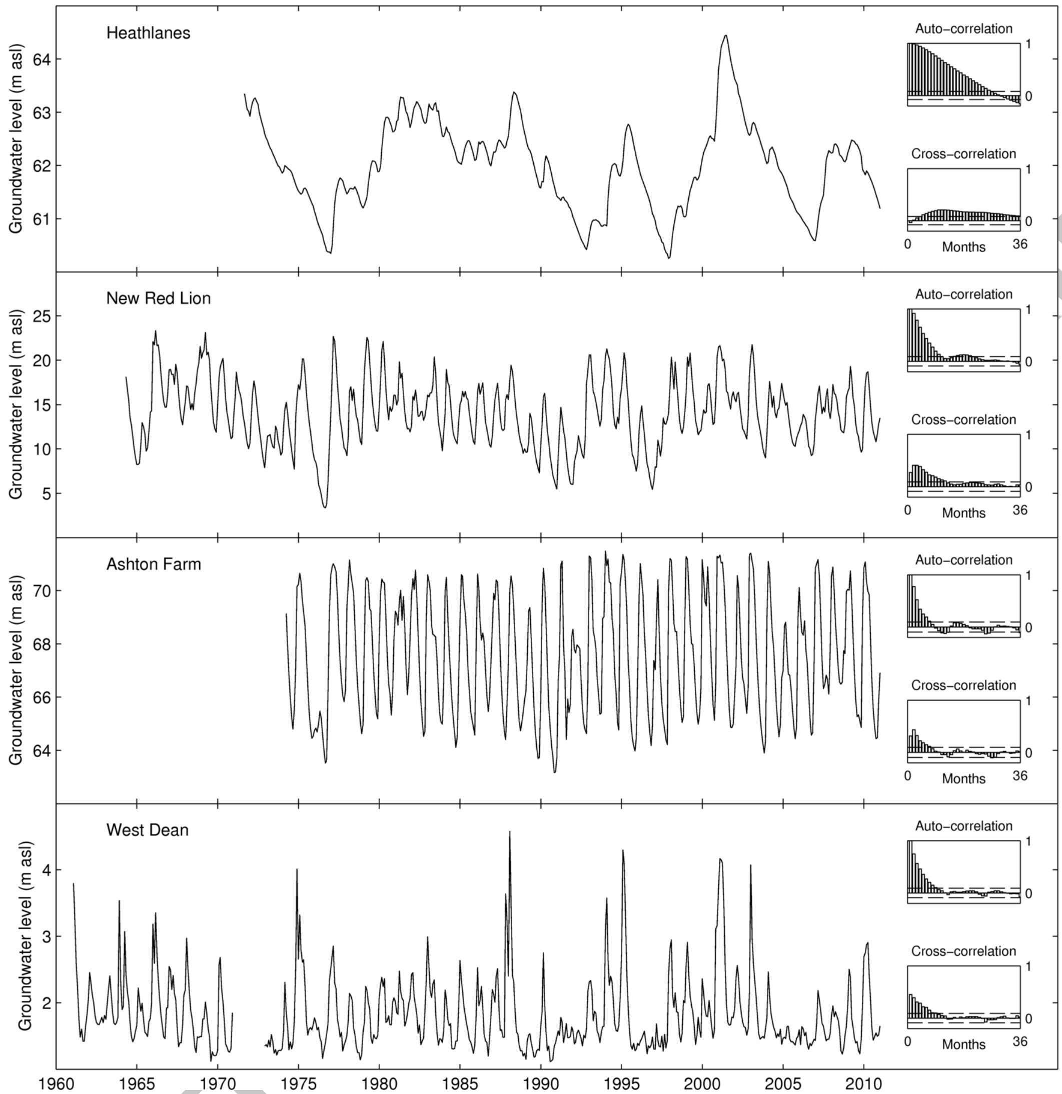


Figure 2: Groundwater level time-series with groundwater level auto-correlation and rainfall-groundwater level cross-correlation plots. Note that the vertical scales vary across the plots.

ACCEPTED MANUSCRIPT

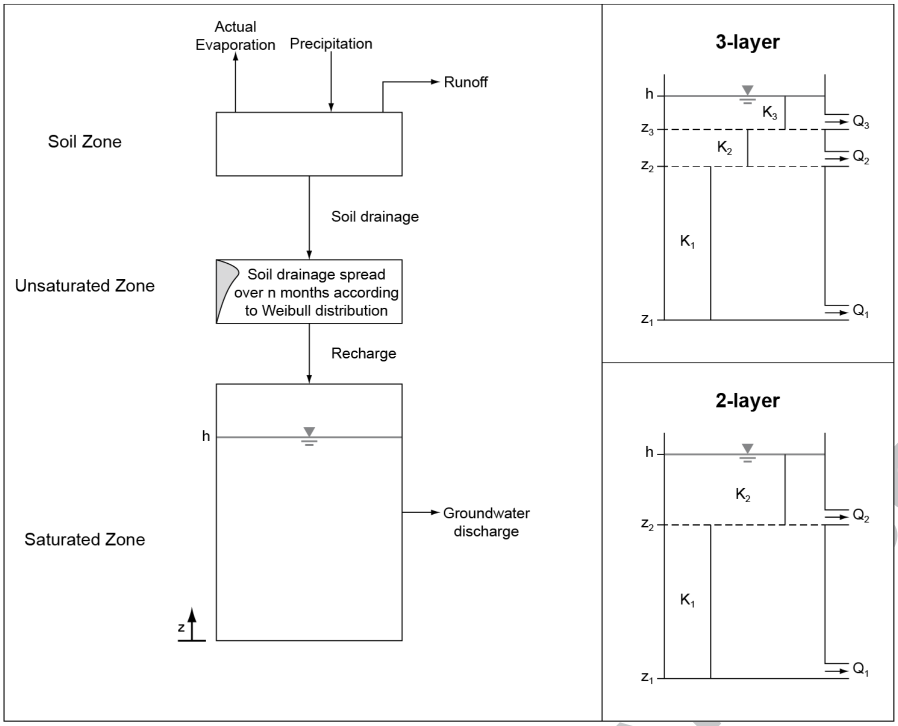
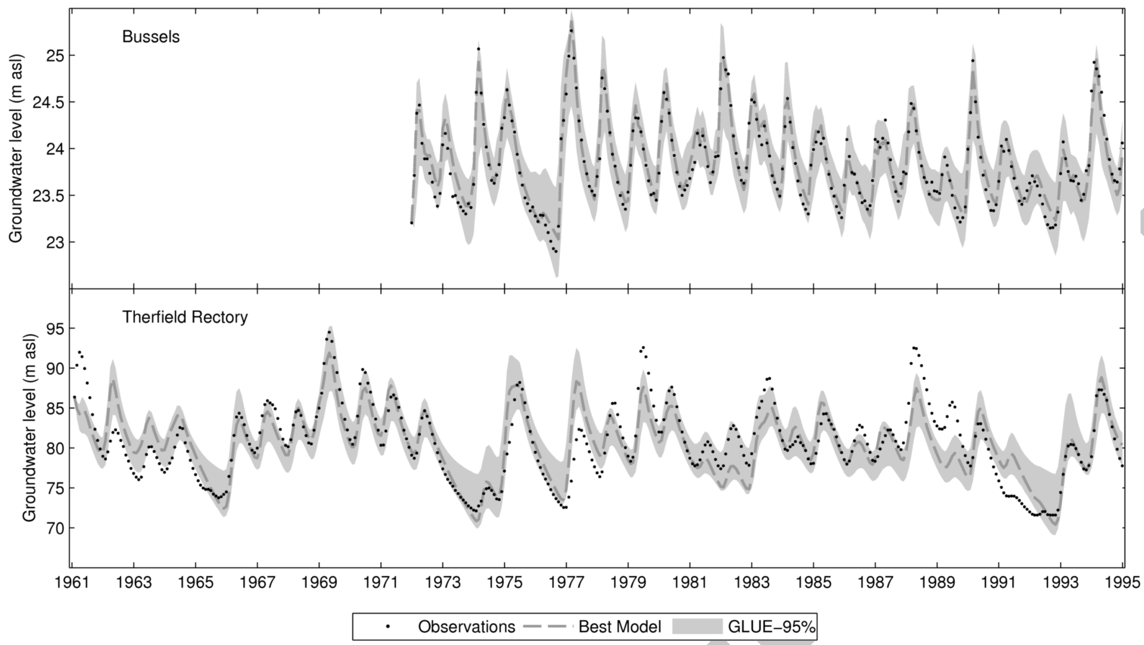


Figure 3: Schematic of generalised AquiMod model structure (left) and different saturated zone component structures used in this study (right) after Mackay et al. (2014).

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

910 

911 *Figure 4: Calibration period simulations and observations for the Bussels and the Therfield Rectory observation*
912 *boreholes.*

913

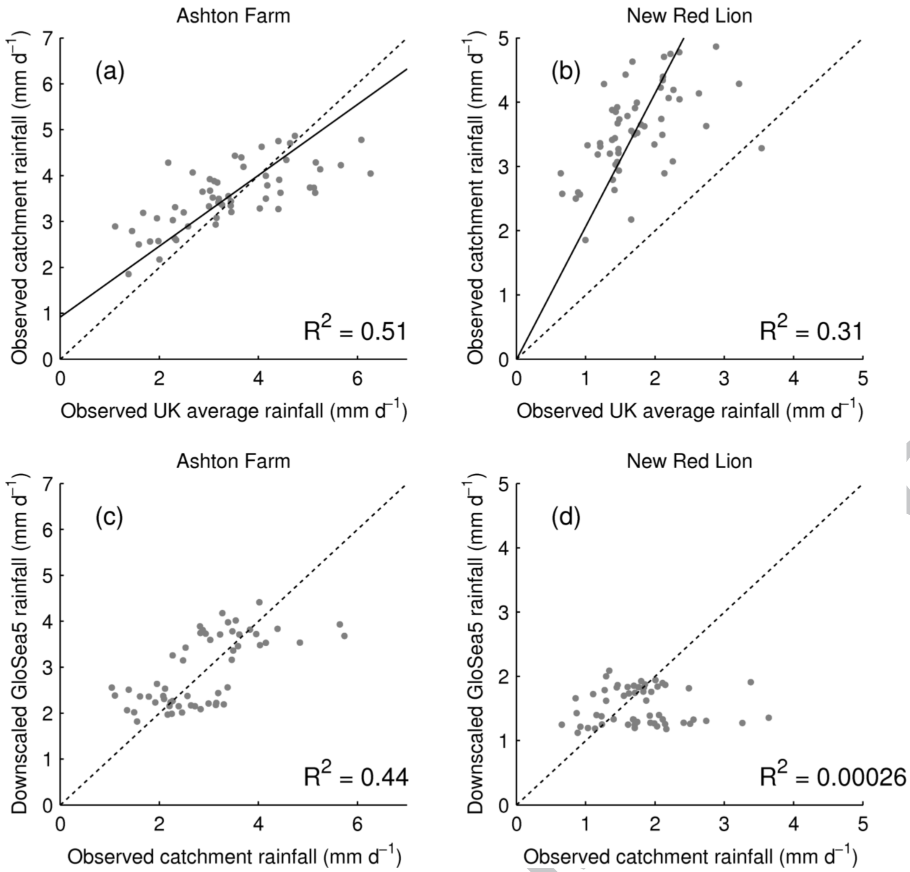


Figure 5: Linear regression models (solid black lines) fitted to downscale seasonal rainfall from UK average to catchment scale for the Ashton Farm (a) and New Red Lion (b) observation boreholes. The resulting correlation between the downscaled GloSea5 rainfall forecasts and the observed catchment rainfall is also shown for the Ashton Farm (c) and New Red Lion (d) observation boreholes.

ACCEPTED MANUSCRIPT

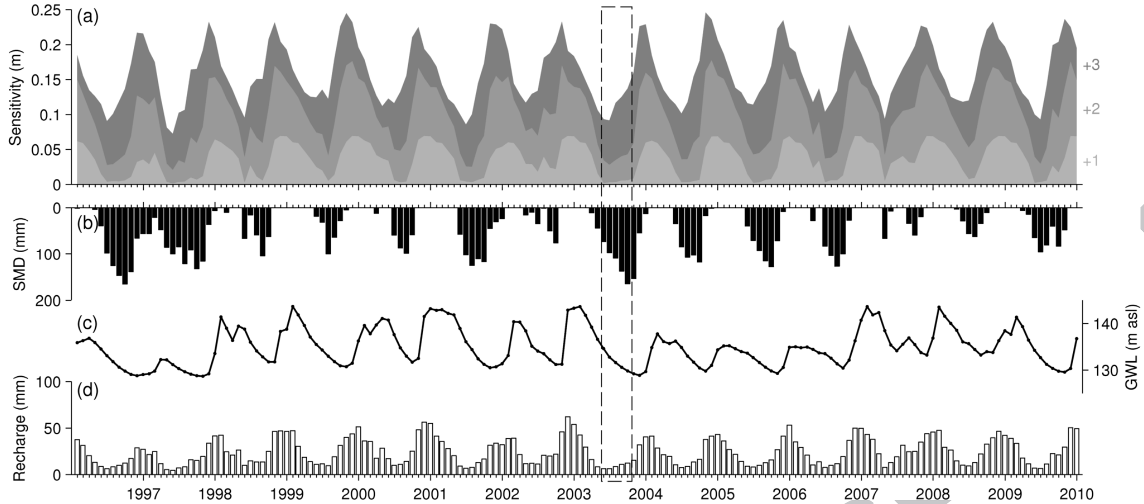


Figure 6: Calculated monthly sensitivity to climate inputs for one, two and three month forecasts (a) ; Soil moisture deficit initial condition (b); Groundwater level initial condition (c); and mean monthly simulated recharge (d).

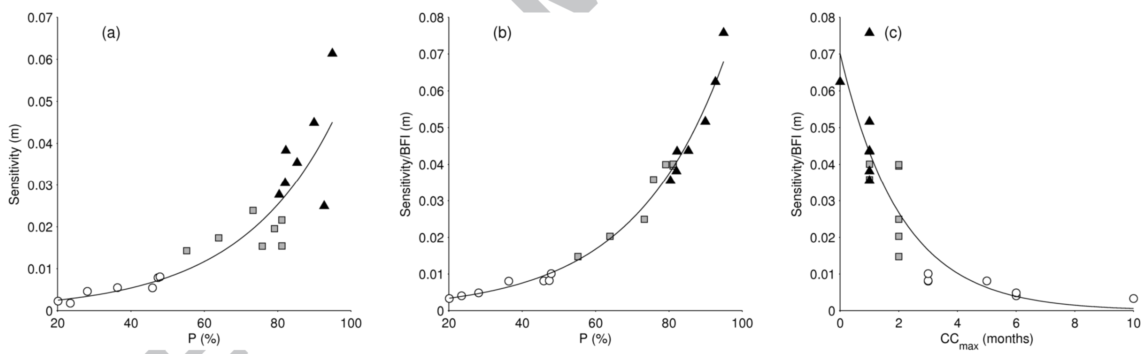


Figure 7: Model response characteristic, P, against the derived model sensitivity (a); P against the sensitivity normalised with respect to the BFI (b); and the peak lead lag correlation between observed rainfall and de-seasonalised groundwater levels, CC_{max}, against the sensitivity normalised with respect to the BFI (c) for the 21 catchment models. All data points are arranged into slowly (circles), moderately (squares) and quickly (triangles) responding catchments.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

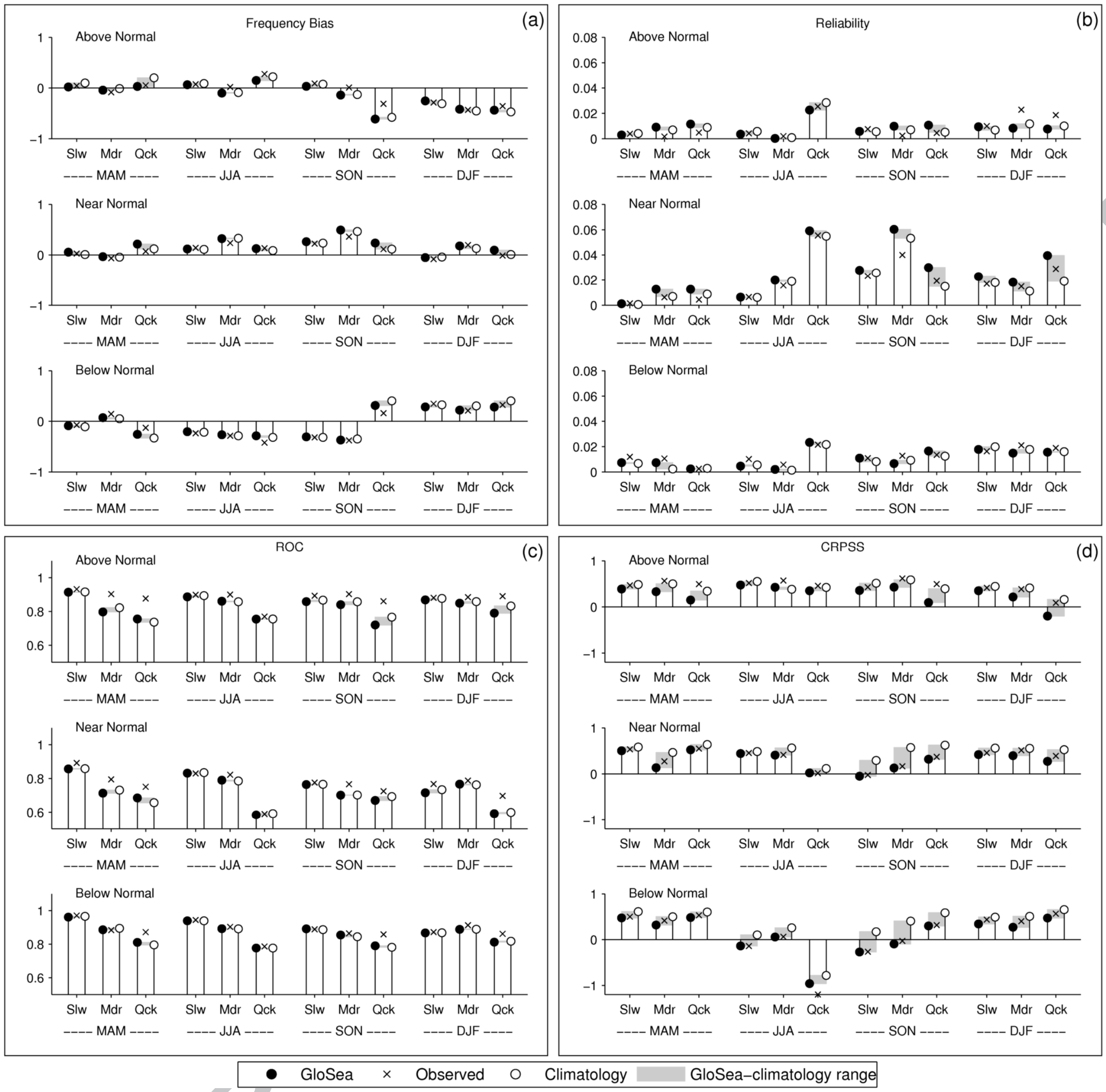


Figure 8: Frequency bias (a), reliability (b), ROC (c), and CRPSS (d) metrics calculated from the reforecasts.

ACCEPTED MANUSCRIPT

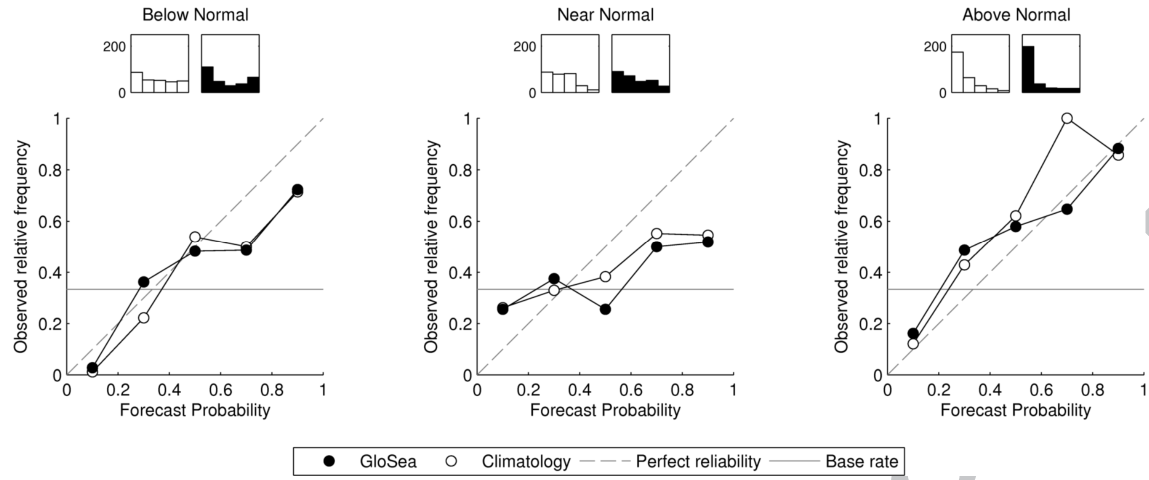


Figure 9: Reliability diagrams for the quickly responding catchments. The histograms denote the sample sizes for each point on the reliability curves.

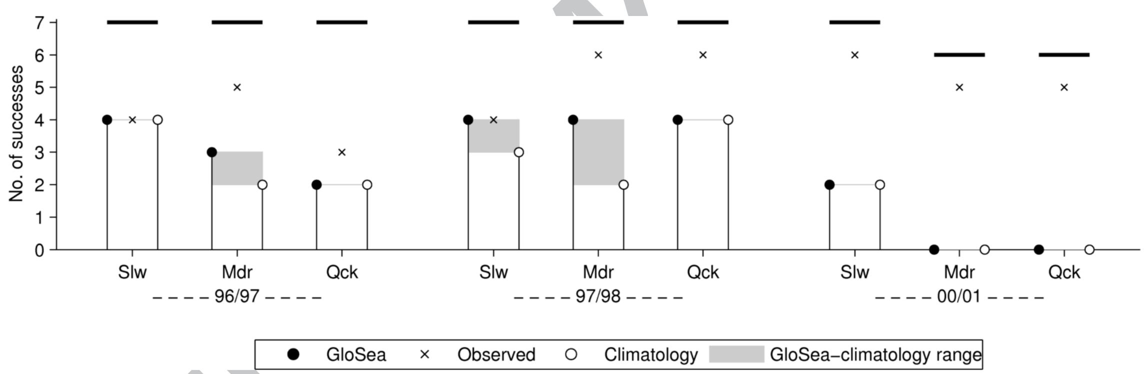


Figure 10: Number of successful forecasts for three events. The solid black lines indicate the total number of catchments with available observation data.

ACCEPTED MANUSCRIPT

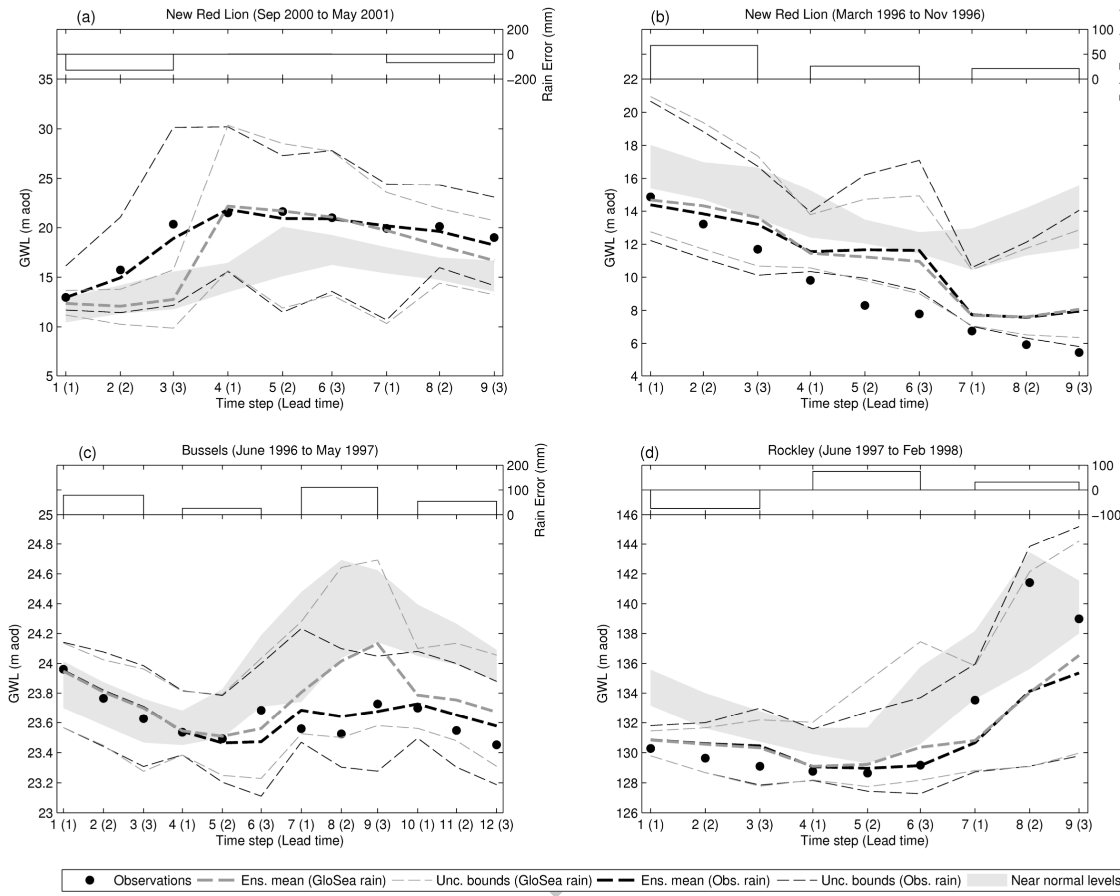


Figure 11: Comparison of the reforecasts using downscaled GloSea and observed rainfall inputs for different time periods.

ACCEPTED MANUSCRIPT

2. Tables

Table 1: List of AquiMod model parameters and calibration ranges.

Module	Parameter (units)	Description	Typical calibration range
Soil	Δx (km)	Representative aquifer length	Fixed as distance between observation borehole and river discharging groundwater.
	BFI (-)	Baseflow index	Taken from Marsh and Hannaford (2008).
	FC (-)	Field capacity of the soil	Taken from Boorman et al. (1995).
	WP (-)	Wilting point of the soil	Taken from Boorman et al. (1995).
	Zr (mm)	Maximum rooting depth of vegetation	100 – 3000
	p (-)	Depletion factor of vegetation	0 – 1
Unsaturated Zone	n (-)	Maximum number of time-steps taken for soil drainage to reach the groundwater	Set based on cross-correlation analysis between rainfall and groundwater levels.
	k (-)	Weibull shape parameter	1 – 7
	λ (-)	Weibull scale parameter	1 – 12
Saturated Zone	K_i (m d^{-1})	Hydraulic conductivity for layer i	0.01 – 100
	S (%)	Aquifer storage coefficient	0.1 – 20
	Z_i (m asl)	Outlet elevation for layer i	Deep outlet set to the known bottom elevation of aquifer.

ACCEPTED MANUSCRIPT

Remaining outlet elevations set after preliminary calibration runs.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

981 *Table 2: List of 21 observation boreholes with the number of behavioural models (n), the efficiency of the most*
982 *efficient model (NSE_{max}) and the containment ratio using the GLUE 95% confidence bounds (CR).*

Observation borehole	**Aquifer**	**n**	**NSE_{max}**	**CR**
Ashton Farm	Chalk	2155	0.89	94.4
Aylesby	Chalk	2470	0.82	96.9
Chilgrove House	Chalk	2125	0.91	97.8
Clanville Lodge	Chalk	2025	0.84	89.0
Dalton Holme	Chalk	2000	0.81	82.6
Grimes Graves	Chalk	1960	0.86	88.9
Little Bucket Farm	Chalk	2305	0.90	85.7
Rockley	Chalk	1835	0.88	94.1
Stonor Park	Chalk	2430	0.78	65.3
Therfield Rectory	Chalk	1915	0.71	68.9
Washpit Farm	Chalk	1910	0.91	96.3
Well House Inn	Chalk	1850	0.73	68.1
West Dean	Chalk	2210	0.83	92.2
West Woodyates Manor	Chalk	1780	0.86	84.8
New Red Lion	Jurassic Limestone	2155	0.74	77.0
Lower Barn Cottage	Lower Greensand	2120	0.81	79.5
Swan House	Magnesian Limestone	1960	0.86	89.6
Bussels	Permo-Triassic Sandstone	2090	0.94	97.5
Furness Abbey	Permo-Triassic Sandstone	2055	0.75	72.7
Heathlanes	Permo-Triassic Sandstone	2095	0.87	87.9
Skirwith	Permo-Triassic Sandstone	2390	0.83	87.6

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Highlights

- We forecast groundwater levels 3 months into the future for 21 boreholes in the UK.
- We use GloSea5 seasonal rainfall forecasts to drive a conceptual groundwater model.
- The forecasts consistently show more skill than a persistence forecasting approach.
- The forecasts are not able to capture extreme groundwater level events.
- Sensitivity to (skill derived from) rainfall forecasts is highly site specific.

ACCEPTED MANUSCRIPT