*A percentile approach to evaluate simulated groundwater levels and frequencies in a Chalk catchment in Southwest England. The authors use the VarKarst model to predict the variation of discharge and groundwater levels in a catchment in England. The topic is relevant to the journal and the work is timely given a growing interest in the forecasting and characterisation of floods and droughts. It would be very valuable to have a discharge/groundwater level model that gives reliable predictions even when the calibration datasets are small. The paper is suitably concise and the description is generally clear. However, I have a number of serious concerns about the focus of the manuscript and the calculations within it. I am unable to recommend the manuscript for publication unless these concerns are addressed.*

We thank the referee for her /his evaluation detailed comments. We appreciate the concerns about our claims about the novelty of the approach. We hope that we can clarify all of the referees concerns in the following response.

1. *In the title and introduction, the authors promote their 'percentile approach' to assessing the performance of the models as the main novelty in the manuscript. I am afraid that I am not persuaded that the percentile approach is novel enough to merit publication in itself. The approach is a comparison between the realised percentiles of the observed and modelled discharge/groundwater levels. It appears to be exactly equivalent to the standard statistical procedure of comparing the distributions of two variables in terms of their realized quantiles. This is a very well used approach, as evidenced by the Wikipedia page describing the QQ plots that result: https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot*

The referee is right. The novelty of our research is the application of a process-based model instead of a statistical distribution function. We admittedly created a wrong perception by the choice of our title. The revised manuscript will the titled with

"Process-based modelling to evaluate simulated groundwater levels and frequencies in a Chalk catchment in Southwest England "

In addition, we will provide more reference to the work of others that applied quantile-quantile approaches in groundwater frequency analysis (see also our reply to Other Comment 1 of the other referee).

2. *Furthermore, I am not convinced that the percentiles used by the authors are a good indicator of the performance of a discharge/groundwater level model. The authors are only confirming that the complete set of modelled values are similar to the complete set of observed values. They are not confirming that the groundwater levels are predicted at the correct time. In terms of the authors' percentile criterion, there would be no penalty for a model that predicts a flood at the time of a drought but compensates by predicting a drought at the time of a flood. For these reasons, I believe that a substantial change of theme of the manuscript is required.*

We believe this is a misunderstanding, in this study, we used the percentile approach only for our evaluation. The calibration and evaluation of our model was carried out with continuous flow and water level observations. The error function KGE that we used for comparing model simulations and

observations explicitly evaluates the correctness of timing by using the linear correlation coefficient *r* as one of its three components (see also our response to general comment 5).

We believe that evaluating the performance of a discharge/groundwater level against percentiles (alongside other metrics such as KGE) is a valuable diagnostic tool. In this case, by evaluating the performance of VarKarst against percentiles we gain insights of which aspects of the flow range (i.e. low flows, high flows) VarKarst is able to reproduce well/poorly. These insights you would not gain with a KGE score and is particularly important when you want to investigate the impacts of expected future climate change on the flow regime as was done in this study.

In the new version of the manuscript we will clarify the model calibration and evaluation to avoid further misunderstanding.

3. *The theme that most interests me in the manuscript is the quest to "balance model complexity and data availability" referred to in the Abstract. If the authors could demonstrate that they have achieved this for their study area then they would have a very valuable paper. However, I believe that much more evidence of this is required. The authors calibrate the 13 parameters of the VarKarst model using data from three boreholes and one timeseries of discharge data. In any such modelling exercise I am concerned whether the parameters maintain their physical meaning and whether the internal processes in the model (e.g. the soil and epikarst modules) are reflecting reality. It is entirely possible that the model is acting as a 'black box' where the large number of parameters are giving it the flexibility to reproduce almost any relationship between the input and output data with which it is presented. If this were the case, it is unlikely that the model would perform well if the characteristics of the input data were to change (e.g. under climate change).*

We understand the concern of the referee. Indeed, models with more than 5-6 parameters are often regarded to end up in equifinality (Jakeman and Hornberger, 1993; Wheater et al., 1986; Ye et al., 1997), i.e. their parameters lose their identifiability (Beven, 2006; Wagener et al., 2002). In such cases, the model can be regarded as a "black box" with rather limited prediction skills as correctly stated by the referee.

In order to reflect the complexity of karst hydrology, 5-6 parameter are often not enough to include all relevant processes in a simulation model. For that reason, recent research took advantage of auxiliary data, such as water quality data or tracer experiments (Hartmann et al., 2013; Oehlmann et al., 2015). These studies showed that adding such information allows identifying the necessary model parameters, therefore enabling the model to reflect the relevant processes.

In this study, we followed this idea and used a combination of groundwater level observations at three locations and discharge observations to obtain enough information to estimate our model parameters. Applying the Shuffled Complex Evolution Metropolis algorithm (also see our response to general comment 5 below) and step wise increasing the calibration data (only discharge, only groundwater, all together), we show that discharge alone, as well as groundwater alone, do not provide enough information to identify all of our model parameters (Fig 5 in the manuscript) as the posteriors of some of the model parameters remain close to a uniform distribution.

Using all information, observed discharge and observations of three groundwater levels, all model parameters are identifiable. I.e., their posteriors strongly differ from a uniform distribution (blue lines in Fig 5), which is in accordance with preceding research that showed that a combination of groundwater and discharge observations can parameter uncertainty (Kuczera and Mroczkowski, 1998). Furthermore, the split-sample test indicates a stable performance of groundwater simulations (Table 3, also see our response to general comment 6). We therefore believe that there is enough indication that the model reproduces the system behaviour satisfactory and that it can be used for prediction.

In the revised manuscript, we will add this clarification to the discussion.

4. *One piece of evidence of the model reflecting reality rather than acting as a black box would be clearly identifiable parameter values. The authors are therefore quite correct to explore the identifabilty of the parameters using the MCMC approach. Their results (Figure 5) indicate that for their final calibration that the parameters are almost perfectly identifiable. Given the short duration, high seasonality and marked temporal correlation amongst the input data I find this surprising. Indeed when (Schoups and Vrugt, 2010) calibrated their similarly complex river models using an MCMC approach many of the parameter values could not be identified. This makes me question the authors' implementation of the MCMC approach.*

Thanks for this critical comment. We thoroughly studied the work of Schoups and Vrugt (2010) in relation to our results. Using a hydrological model with seven parameters combined with an error model with 4-5 parameters their calibration problem is indeed similar to the one we present in our study. But there is one important difference: They only use discharge observations for model calibration. As found by many preceding studies (see also our response to general comment 2) simulation models with more than 5-6 parameters typically result in increased parameter uncertainty, which Schoups and Vrugt (2010) also found in their study. Using only discharge information, our study would have resulted in similar problems (green lines in Fig 5). However, the combined use of discharge observations and the observations of three groundwater wells resulted in increased parameter identifiability, as we could also show in Fig 5 (blue lines). Therefore, our results do not contradict with Schoups and Vrugt (2010) but they rather show that there are ways to reduce parameter uncertainty by auxiliary data (see also our response to general comment 2).

In the revised manuscript, we will put more emphasis on the description of these multiple data sets for parameters estimation and refer to the comparison with Schoups and Vrugt (2010) in the discussion.

5. *Within a MCMC algorithm, a huge number of different sets of parameter values are compared. Those sets that are consistent with the observed data are included in the Markov chain whereas other parameter sets are discarded. These comparisons are normally made by calculating the likelihood function for the different parameter sets (e.g. Schoups & Vrugt, 2010). It is possible to use the calculated likelihoods or probabilities to determine which parameters sets are good enough to be included in the Markov chain. Thus, the inclusion or exclusion of a parameter set is decided by an objective criterion that is consistent with statistical theory.*

*It appears that the authors have compared different parameter sets in terms of their KGE score. This concerns me because it is not clear to me how to decide what magnitude of difference between KGE scores signifies that one set of parameters is not good enough to be included. A threshold on the KGE scores could be set arbitrarily but then the realised distributions of the parameters become meaningless. The apparent identifiability of the parameters could be changed by a simple and arbitrary tweak of this threshold.*
*Therefore, the authors must give more detail about the comparison function they included in the MCMC algorithm and demonstrate how it leads to objective estimates of the posterior distributions of the parameters.*

The Shuffled Complex Evolution Metropolis algorithm (SCEM, Vrugt et al., 2003) that we used in our study is based on the Metropolis-Hastings algorithm (Hastings, 1970; Metropolis et al., 1953) and the Shuffled Complex Evolution algorithm (Duan et al., 1992). The Metropolis-Hastings algorithm uses a formal likelihood measure, i.e. an objective criterion that is consistent with statistical theory, and calculates the ratio of the posterior probability densities of a "candidate" parameter set that is drawn from a proposal distribution and a given parameter set. If this ratio is larger or equal than a number randomly drawn from a uniform distribution between 0 and 1, the "candidate" parameter set is accepted. This procedure is repeated for a large number of iterations. If the proposal distribution is properly chosen, the Markov Chain will rapidly explore the parameter space and it will converge to the target distribution of interest (Vrugt et al., 2003).

In the SCEM algorithm, "candidate" parameter sets are drawn from a self-adapting proposal distribution for each of a predefined number of clusters. Again a random number [0,1] is used to accept or discard "candidate" parameter sets. In our study, we use the Kling-Gupta efficiency KGE (Gupta et al., 2009) as objective function, which can be regarded as an informal likelihood measure (Smith et al., 2008). To decide whether to accept or discard a parameters set, we compare the KGEs of the "candidate" and the given parameter sets. Such procedure was already applied in various studies (Blasone et al., 2008; Engeland et al., 2005; McMillan and Clark, 2009) and is possible if the error functions monotonically increasing with improved performance. We achieved this in the SCEM algorithm by defining $KGE_{SCEM}$ as

$$KGE_{SCEM} = -\sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2}$$

$$\alpha = \frac{\sigma_S}{\sigma_O};\ \beta = \frac{\mu_S}{\mu_O}$$

With $r$ as the linear correlation coefficient between simulations and observations, and $\sigma_S$, $\sigma_O$ and $\mu_S$, $\mu_O$ as the means and standard deviations of simulations and observations, respectively.

As stated correctly by the referee, the shape of the posteriors is dependent on the error function and using another likelihood measure, formal or informal, may have resulted in different shapes of the posteriors. However, applying SCEM with KGE in our stepwise procedure we are mostly interested in the relative differences of the posteriors and we can clearly see how some of posteriors translate from a uniform distribution to a well-defined peak when more information is added (see also our response to general comment 4). These results combined with the acceptable multi-objective performance of the model during calibration and validation (see also our response to general comment 6), and the realistic parameters that we finally found (see discussion of parameter values in

subsection 5.3) makes us confident that the model reproduces the relevant features of our studied system.

In the revised manuscript, we will provide this more detailed elaboration in the methods section and discuss the consequences of using an informal likelihood measure in the revised discussion.

6. *I'd also like clarification about how the authors decided that their validation results were sufficiently good to conclude that "the model provides robust simulations of discharge and groundwater levels". The authors state that the difference between the calibration and validation KGE scores are small. For each data source, the validation results are worse than the calibration results. Might this indicate that the model is too complex? How big a difference between validation and calibration results would have been required for the authors to conclude that the model had been ineffective?*

Split-sample tests are a common and necessary tool to evaluate the prediction performance of a simulation model (Klemeš, 1986). If the model is compared to a validation period, i.e. a time series of observations that was not used for parameters estimation, a decrease of performance has to be expected because there is always a tendency to compensate for model structural limitations and observational uncertainties during the calibration. If a model contains too many degrees of freedom (model parameters), there is a risk that calibration may overcome all these limitations and uncertainties although the model is a poor choice for the studied system. A split sample-test would indicate such failure by a strong decrease of performance during the validation period.

As correctly questioned by the referee the threshold, from which a decrease of performance is not acceptable anymore, is subject to the individual case of application and the opinion of the modeller. In our case we obtained a decrease of performance from -11% (groundwater prediction) to -21% (discharge prediction). Such ranges are comparable with split-sample tests found by other studies (-4% to -14% by Parajka et al., 2007; -5% to -24% by Perrin et al., 2001). The lower decrease in performance that we found for the simulation of groundwater levels also indicates more stable prediction performance for the groundwater simulations that we later use for our example application with the simplified climate scenarios.

We will add these aspects to the discussion of results of the split sample test in the revised manuscript.

7. *There is a great deal of seasonality in the groundwater levels. Can we be sure that the model is going beyond these seasonal trends? Could a simple annual periodic function have given similarly good results and better managed the trade-off between model complexity and data availability?*

Yes, a simple annual periodic function may be able to reproduce observed variability of groundwater levels to some degree. However, such a function would not be more than a black box model and it would not be straight forward using it to assess the impact of climatic changes on groundwater levels. The structure of the VarKarst model takes into account the particulates of karst hydrology (see also our response to specific comment 1). We believe that our analysis and evaluation provides some

indication that it is also able to reflect the observed processes at our Chalk study site, therefore making it a useful tool to explore the impact of climate changes on groundwater level dynamics.

We will clarify this in the introduction of the revised manuscript.

Specific comments:

1. *The introduction provides a clear description of the hydrogeological system with the appropriate level of description and ample references for anyone who wants to delve further (the same can also be said of section 2). More detail could be provided in the paragraph which describes the importance of the work in this study. I appreciate that the authors have made the Methodology section concise by referring to previous papers. However, I think they could give a clearer overview of the VarKarst model whilst leaving the details to the other papers. What do the 15 model compartments correspond to? Are they situated along some sort of gradient in the catchment? If so, is it possible to use knowledge of the hydrogeological system to determine the compartment in which each borehole is situated? What do they mean when they say that the spatial variability of the soil, epikarst and groundwater systems are expressed as a Pareto function? - What characteristics of these systems are the authors referring to? - Are these characteristics sampled from a Pareto distribution or do they decay according to a Pareto function?*

We thank the referee for these helpful suggestions. We will highlight the scope and importance of this work in the introduction of the revised manuscript. In addition, we will add a more detailed description of the VarKarst model and the meaning of its individual components (including the distributed model compartments) to the appendix.

2. *Equation (1). Ensure that all symbols in all equations are defined. Use a multiplication sign rather than '*'.*

We agree. We will improve our manuscript by eliminating all stylistic flaws.

3. *Section 3.3 – Give more detail about the implementation of the MCMC algorithm to address my concerns above. In particular, explicitly state the function used to decide whether a parameter set is accepted or rejected and explain how these lead to objective and representative samples of the posterior distributions.*

Please see our response to general comment 4.

4. *Section 3.4 The authors state that their percentile approach was motivated by standardised groundwater and precipitation indices. Seasonality is often removed from standardised indices. Did the authors consider removing seasonality from their simulations before assessing them?*

During the period of model development and calibration, we considered calibrating directly to the flow percentiles, i.e. removing seasonality. However, removing the temporal information from the time series would have reduced the information content of the data and would have resulted in increased parameter uncertainty (see our response to general comment 2 and 3) with a lower prediction performance of the model.

We will add this information to the methodology of the revised manuscript.

5. *Equation (2) Write words such as mean in standard font rather than italics. State which variable you are summing over. The authors calculated the 5th percentile at a yearly time scale using 10 years of data. Does this mean they attempted to determine the 5th percentile from only 10 observations of yearly data?*

We will improve the elaborations on Eq. 2. The percentiles were derived from the daily data of the calibration period (2008-2012). We then compared the average sum of days exceeding the respective percentile in the respective time scale. We will also clarify this in the revised manuscript.

6. *Section 3.5: I am not sure that using nine climate scenarios is sufficient to assess the uncertainty in the effect of climate change on groundwater levels.*

The purpose of the simple climate scenarios was to provide an application example of the new methodology, which is rather hypothetical considering the large uncertainties of current climate projections. We believe that our 9 realisations are sufficient to show that different possible future changes have a non-linear impact on groundwater level frequencies.

In the revised manuscript we will clarify the simple exemplary characteristic of this application of our approach.

7. *Section 4.2. The poor performance of the model when groundwater levels are large could be because the authors are using an objective function that is suited to Normally distributed variables but the distribution of groundwater levels are skewed. Have the authors tried an objective function that is more suited to skewed data?*

The objective function (KGE) is applied to simulated time series of groundwater levels. It was chosen by trial and error comparing the simulation performances during calibration and validation obtained different objective functions (RMSE and other). We found that we obtain the most robust results with the KGE.

We will mention this trial and error procedure in the Methods sections of the revised manuscript.

**References**


Beven, K. J.: A manifesto for the equifinality thesis, J. Hydrol., 320(1–2), 18–36, 2006.

Blasone, R. S., Vrugt, J. A., Madsen, H., Rosbjerg, D., Robinson, B. A. and Zyvoloski, G. A.: Generalized likelihood uncertainty estimation (GLUE) using adaptive Markov Chain Monte Carlo sampling, Adv. Water Resour., 31(4), 630–648, doi:10.1016/j.advwatres.2007.12.003, 2008.

Duan, Q. Y., Sorooshian, S. and Gupta, H. V: Effective and Efficient Global Optimization for Conceptual Rainfall-Runoff Models, Water Resour. Res., 28(4), 1015–1031, 1992.

Engeland, K., Xu, C.-Y. and Gottschalk, L.: Assessing uncertainties in a conceptual water balance model using Bayesian methodology / Estimation bayÃ©sienne des incertitudes au sein dâ€™une modÃ©lisation conceptuelle de bilan hydrologique, Hydrol. Sci. J., 50(1), 45–63 [online] Available from: http://www.informaworld.com/10.1623/hysj.50.1.45.56334, 2005.

Gupta, H. V, Kling, H., Yilmaz, K. K. and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, J. Hydrol., 377(1–2), 80–91, doi:10.1016/j.jhydrol.2009.08.003, 2009.

Hartmann, A., Barberá, J. A., Lange, J., Andreo, B. and Weiler, M.: Progress in the hydrologic simulation of time variant recharge areas of karst systems – Exemplified at a karst spring in Southern Spain, Adv. Water Resour., 54, 149–160, doi:10.1016/j.advwatres.2013.01.010, 2013.

Hastings, W. K.: Monte Carlo Sampling Methods Using Markov Chains and Their Applications, Biometrika, 57(1), 97–109 [online] Available from: http://www.jstor.org/stable/2334940, 1970.

Jakeman, A. J. and Hornberger, G. M.: How much complexity is warranted in a rainfall-runoff model?, Water Resour. Res., 29, 2637–2649, 1993.

Klemeš, V.: Dilettantism in Hydrology: Transition or Destiny, Water Resour. Res., 22(9), 177S–188S, 1986.

Kuczera, G. and Mroczkowski, M.: Assessment of hydrologic parameter uncertainty and the worth of multiresponse data, Water Resour. Res., 34(6), 1481–1489, 1998.

McMillan, H. and Clark, M.: Rainfall-runoff model calibration using informal likelihood measures within a Markov chain Monte Carlo sampling scheme, Water Resour. Res., 45(4), 1–12, doi:10.1029/2008WR007288, 2009.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E.: Equation of State Calculations by Fast Computing Machines, J. Chem. Phys., 21(6), 1087, doi:10.1063/1.1699114, 1953.

Oehlmann, S., Geyer, T., Licha, T. and Sauter, M.: Reducing the ambiguity of karst aquifer models by pattern matching of flow and transport on catchment scale, Hydrol. Earth Syst. Sci., 19(2), 893–912, doi:10.5194/hess-19-893-2015, 2015.

Parajka, J., Merz, R. and Blöschl, G.: Uncertainty and multiple objective calibration in regional water balance modelling: case study in 320 Austrian catchments, Hydrol. Process., 21(4), 435–446, doi:10.1002/hyp.6253, 2007.

Perrin, C., Michel, C. and Andréassian, V.: Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments, J. Hydrol., 241, 275–301, 2001.

Schoups, G. and Vrugt, J. A.: A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors, Water Resour. Res.,

46(10), 1–17, doi:10.1029/2009WR008933, 2010.

Smith, P., Beven, K. J. and Tawn, J. A.: Informal likelihood measures in model assessment: Theoretic development and investigation, Adv. Water Resour., 31(8), 1087–1100, doi:10.1016/j.advwatres.2008.04.012, 2008.

Vrugt, J. A., Gupta, H. V, Bouten, W. and Sorooshian, S.: A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters, Water Resour. Res., 39(8), 18, 2003.

Wagener, T., Lees, M. J. and Wheater, H. S.: A toolkit for the development and application of parsimonious hydrological models, Math. Model. large watershed Hydrol., 1, 87–136, 2002.

Wheater, H. S., Bishop, K. H. and Beck, M. B.: The identification of conceptual hydrological models for surface water acidification, Hydrol. Process., 1(1), 89–109, doi:10.1002/hyp.3360010109, 1986.

Ye, W., Bates, B. C., Viney, N. R., Sivapalan, M. and Jakeman, A. J.: Performance of conceptual rainfall-runoff models in low-yielding ephemeral catchments, Water Resour. Res., 33(1), 153–166, doi:10.1029/96wr02840, 1997.