Nat. Hazards Earth Syst. Sci. Discuss.,
doi:10.5194/nhess-2016-386-RC2, 2017

Natural Hazards and Earth System Sciences
Discussions

Interactive comment on "A percentile approach to evaluate simulated groundwater levels and frequencies in a Chalk catchment in Southwest England" *by* Simon Brenner et al.

Anonymous Referee #2

Received and published: 16 February 2017

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.

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

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.

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).

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.

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.

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? 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?

Specific comments:

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?

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

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.

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?

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?

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.

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?

References

Schoups, G., Vrugt, J.A., 2010. A formal likelihood function for parameter and predic-

tive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. Water Resources Research, 46, W10531.

Interactive comment on Nat. Hazards Earth Syst. Sci. Discuss., doi:10.5194/nhess-2016-386, 2016.