Our responses are in Times New Roman following the individual comments in courier.

K. J. Beven (Referee)

k.beven@lancaster.ac.uk

Received and published: 14 April 2013

This is a useful demonstration of the equivalence of GLUE and ABC results, and of the utility of PMC sampling as a way of increasing the efficiency of sampling an ensemble of behavioural models. But it leaves a number of questions unaddressed.

Response: The main point of this paper is simply to demonstrate that if all observations are used as summary metrics GLUE is a special variant of Approximate Bayesian Computation. No doubt that we left a large number of other questions unaddressed. This is a largely unexplored idea that has great prospects but also brings up new problems to solve.

1. In what sense is ABC more generic than GLUE. In fact GLUE, even as set out in 1992 has a much wider range of options (and has been used since with formal likelihood functions). So why is ABC not considered as a special case of GLUE rather than the other way round (ABC is not actually even formally Bayesian in that it makes no use of Bayes equation)?

Response: The ABC procedure has its roots within likelihood-free inference proposed by Diggle and Gratton in 1984. This latter approach has a sound statistical underpinning and has become a common theme within the mathematical/statistical literature. So, the ABC work has a much better foundation, and its foundation has been laid years before GLUE has come out.

With respect to ABC not being Bayesian. The results of ABC converge theoretically to those of formal Bayesian approaches if a sufficient set of statistics is used. Information theory can help to determine which statistics to use, and when sufficiency is achieved. The advantage of ABC is however, that it does not require explicit definition of a likelihood function. The user is free to select which statistic, rooted in hydrologic theory, is deemed appropriate. The methodology has a stronger diagnostic power as previously illustrated in Vrugt and Sadegh (2013).

2. After all the past criticism of GLUE "lumping all uncertainty into parameter distributions", why is the implicit/explicit

error treatment issue totally ignored here, to the extent that formal likelihood results are presented (wrongly) only in terms of the posterior parameter predictions.

Response: We are a bit confused here. How does GLUE handle different error sources? To the best of our knowledge, the approach that Beven has advocated is to use limits of acceptability. The main scope of this paper is to show that this is a special case of ABC. The places where we use DREAM is to simply compare our findings with those of a formal Bayesian approach (with an unavoidably incorrect likelihood function).

3. The authors do not clearly separate the two issues of defining criteria for choosing behavioral models and sampling the resulting model space. As noted below, efficient sampling methods help, but it is the choice of criteria that will control how complex the space is to be searched. It is also possible that the convergence of more efficient sampling methods will fail to identify local areas of behavioural models even given the random steps of MCMC type methods. I agree that efficiency is an issue – but the choice of criteria is much more important.

Response: We appreciate this comment, but the scope of this paper is not which summary metrics to choose, or how to improve sampling efficiency. We simply demonstrated a significant level of agreement between GLUE limits of acceptability and ABC. Our recent work has developed a new sampling method that is 10-20 times faster than current rejection and Population Monte Carlo (PMC) samplers promulgated in the statistical literature. But this is material for a different paper. So is the selection which summary metrics to use. Information theory will determine sufficiency. We have over 110 different metrics that we are currently testing for adequacy – those will be considered in future work.

4. The results reveal that the calibration/validation process is subject to epistemic errors (as discussed for these same data sets by the Beven, 2009 comment on Vrugt et al. 2008). The method of estimating a reasonable range of acceptability used here, reveals something about the errors at least in calibration. But these are not then used in prediction (as formal error should be for the full Bayes approach), the results presented are based only on the posterior parameter distributions. Why not? Surely this is important in deciding whether the resulting ensembles should be considered fit for purpose or not (the authors make no comment as to whether bracketing 68% of the observations is fit for purpose – is it not indicating something about errors in either model or data?).

Response: We could have used the error structures in calibration during the evaluation period, but we purposely decided to propagate parameter uncertainty only. The consistency in performance between the calibration and evaluation period determines whether we derived reasonable distributions.

I conclude that the paper needs major revision in terms of both the presentation and discussion of the results.

Response: We will make the revisions that we deemed are appropriate. In any case we were happy to see the current version of our paper to be cited in a paper from the referee (GLUE-20 year further). The citation was positive!

Some specific comments

4740/9 Abstract. In this paper we introduce an alternative framework, called Approximate Bayesian Computation (ABC) that summarizes the differing viewpoints of formal and informal Bayesian approaches. – what does this sentence mean? ABC does no such thing. Its only claim to resolve the different viewpoint is that for certain toy problems it can be shown to converge asymptotically (but not necessarily quickly) to a formal posterior.

Response: We appreciate this comment of the reviewer, and shall reformulate this sentence. But not only for toy problems will ABC converge to formal Bayesian approaches. It will do so in practice as well, if the correct likelihood function is used within formal Bayes. And this is the crux of the problem in many hydrologic studies, particularly when faced with errors in the forcing data. ABC circumvents this problem by using summary metrics. Those can be defined in such a way that they are insensitive to errors in the precipitation and PET. An interesting study could compare the results of the Generalized Likelihood Function of Schoups and Vrugt (2010) against those with ABC. But even then, the GL will suffer from incomplete treatment of rainfall data errors.

4749/15 The use of such "insufficient statistic" promotes equifinality, and makes it unnecessarily difficult to find the preferred parameter values. - of course, but ABC gives equivalence to all all samples within the threshold of acceptability so does not eliminate equifinality. Indeed, given the types of epistemic error you demonstrate later you would not want to, since otherwise you might be overconditioning based on a particular realization of epistemic error in your calibration data.

Response: Indeed – we would like our posterior parameter distribution to adequately envelop the observation data. But, a minimum ABC requirement is that a set of sufficient statistics is used. Even then, a posterior will be found, that might demonstrate a large simulation uncertainty.

4749/24 The premise behind ABC is that _0 should be a sample from the posterior distribution as long as the distance between the observed and simulated data, : : : is less than some small value. – and how is this different from GLUE then?

Response: This is not different from GLUE. Yet, in ABC much smaller epsilon values are required than in GLUE to demonstrate converge to the exact posterior. That is one of the main reasons to use summary statistics as one cannot expect that in the time-domain the residuals are very small. Obviously we can state that this is similar to GLUE limits of acceptability, yet the choice of epsilon is significantly different.

4751/16 For illustrative purposes we start with the mean of the actual data, - but why not use NSE to make similarity more obvious (and reduce the impact of using such an inefficient statistic)?

Response: We deliberately used the mean of the data. Gupta et al. (2010) has shown that the NSE statistic is a combination of three individual summary metrics, one that measures the mean of the data, one the standard deviation of the data, and the last one that measures the (temporal) correlation between the measured and simulated data. Thus, the NSE consists of three different components. We therefore start with the mean of the data, the first component of the NSE.

4752/13 search. Our sampler therefore adaptively determines the next value of _j ; j > 1 from the cumulative distribution function of the _(Â ˚u) values of the N most recent accepted samples – this might be fine for simple surfaces but would appear to exacerbate the danger of not sampling areas of behavioural models in more complex spaces?

Response: We appreciate this comment. Indeed, this is possible. The PMC sampler is fine for the present studies, but not ideal for complex and high-dimensional search spaces. We see all the problems with existing ABC sampling methods, including a lack of crossover, problems with a threshold acceptance rule and poor updating of the proposal distribution. The scope of this paper is not to introduce a more efficient ABC sampling method. We have developed such procedure by taking advantage of MCMC simulation with DREAM using a continuous rather than discrete kernel to determine whether to accept the candidate point or not. This work is ready for submission and should be published in due course.

4754/8 The distance function specified in Eq. (5) has many elements in common with the triangular, trapezoidal or beta fuzzy-membership functions used in the limits of acceptability approach of GLUE – ???? surely has much more in common with "classic" GLUE thresholding of informal measures, especially since ABC as applied here uses no such weighting function

Response: Point well taken. We will make some editorial changes to the manuscript to more carefully address this commonality. This will not affect the thrust of our paper.

4754/19 Latin Hypercube sampling strategy used in GLUE to find behavioral solutions. – err…LHS has been used in GLUE but not that commonly – and the original 1992 GLUE paper used a nearest neighbor MCMC-type sampler so it is not limited to either uniform or LHS sampling.

Response: We appreciate this comment. Indeed, the work of Blasone et al. (2006) has introduced a MCMC based procedure to sample the behavioral space of solutions. Yet, this and other sampling methods have not become common practice in GLUE. Most GLUE applications reported in the literature resort to relatively simple and inefficient sampling methods.

4755/19 The adaptive updating strategy of _ in PMC not only guarantees a more efficient search strategy than ABC-REJ (GLUE), but also automatically determines the maximum attainable coverage of the discharge observations within the limits of acceptability. – Here you should differentiate between search strategy and defining behavioural simulations. The "true" ensemble of behavioural simulations is not dependent on search strategy – efficiency helps but might also not find the complete sample if there are multiple local areas of behavioural simulations as identified by multiple criteria (also 4758/28 ff).

Response: We appreciate this comment of the reviewer. Simulations demonstrate that the PMC sampler works fine for the dimensionalities and models considered herein. For more complex models, multiple chains are required with crossover and a continuous kernel to adequately search the entire space of solutions. The freezing of epsilon has many elements in common with simulated annealing – which in turn is inspired from MCMC simulation. With a simple rejection algorithm with a single epsilon value, it is extremely difficult to exactly delineate the space of acceptable solutions. The iterative reduction of epsilon, and updating of the sampling distribution helps to better locate the posterior region. This enhances and sampling efficiency.

4756/7 The simulations nicely track the observed data with uncertainty intervals that appear relatively narrow and encompass about 90% of the data. – If I have understood correctly you are plotting only the ensemble of behavioural models in this plot. So you are saying that the implicit handling of model errors in the original GLUE formulation works (at least for this toy example – this was also demonstrated for the Mantovan and Todini toy example in Beven et al., 2008). But surely you cannot just report this, after all the past argument about the "subjectivity" of using an implicit error model (or as some people put it lumping all the error into the parameter distributions) without at least some comment?????

Response: This is a synthetic case study in which the discharge data were corrupted with a heteroscedastic error. So one would assume that the posterior distribution can properly track the data.

4758/10 This provides further support for our claim that the limits of acceptability approach of GLUE can be interpreted as a special case of formal Bayes. – No, surely not – up to now you have shown that ABC produces similar results to GLUE – it is in fact a special case of GLUE since GLUE is more general than the ABC approach you describe.

Response: This is in some sense the chicken and egg problem. Likelihood-free methods have been introduced years before GLUE was proposed. The ABC methodology is becoming a standard framework within the statistical literature, and we therefore use this as our benchmark in part because this method benefits from an appropriate statistical underpinning.

4760/11 The 95% uncertainty ranges derived with both methods encompass about 70% of the discharge observations. This coverage is significantly larger than the approximately 12–17% derived from a classical likelihood function. (also 4761/19) - Whoa!! You are comparing different things here. ABC/GLUE are using an implicit error model, formal Bayes has an explicit model that should be included in the outputs. This would normally cover more than the 70% for ABC/GLUE. You would surely use this if, for example, you were interested in flood forecasting – e.g. Romanowicz et al WRR 2008). Coverage of the parameter uncertainty is a totally inappropriate measure for the formal likelihood.

Response: We appreciate this comment. Indeed, if we would add model error to our simulation then the coverage of formal Bayesian approaches would significantly increase. Yet, we deliberately focus on parameter uncertainty only. There is an advantage to explaining uncertainty with variations in the model parameters, rather than adding a random model error term that reflects the remaining uncertainty. Yet, we will make some changes to the paper to explicitly state that the uncertainty originates from the parameters only.

4761/5 Because of sampling inefficiency the GLUE calculations were terminated after 100 behavioral samples were identified – err, why? It surely does not take much to continue to run while preparing the paper: : :. Again efficiency is helpful, but it is not generally that much of a problem for this type of model.

Response: The GLUE sampling approach is just incredibly inefficient with acceptance rates lower than 1e-4%. Indeed, we could do more simulations, yet our initial results showed that this hardly affects the outcome.

4761/26 – but this issue – and the limitations of rainfall correction have already been discussed in my SERRA 2009 comment on your 2008 paper. It is really a bit naughty not to mention that.

Response: No “bad” or “naughty” intentions here. We can modify the paper to reflect this previous work.

4762/2 This is simply the effect of an increased rainfall intensity during the evaluation period. – how do you know that? Why could it not be some other sort of epistemic error, a consistent increase over such a period would be hydrologically rather strange would it not? Certainly not simply!! The issues of such non-ideal cases are discussed in Beven, 2006, 2010, 2012 and Beven and Smith HESS 2011 – and are even more important for formal likelihoods. Should surely be part of the discussion.

Response: We agree with the reviewer. We cannot prove that this is the consequence of larger rainfall events, and an ensemble that was not trained to fit larger events. Model structural error will surely pay a role as well. We will rephrase this sentence in the revised paper, with due attention to the cited papers.

4764/14 The effective observation error remedies this problem, but the magnitude of this value is typically much larger than the theoretical value of _ to guarantee converge to the true posterior parameter distribution. No, this is totally the wrong

argument. There can be no true parameter distribution for this type of non-ideal problem, only for toy problems (and if you believe you have a toy problem then why not use a formal likelihood approach). The limits you are defining are related to the all the observational uncertainties you mentioned earlier (and you should really be also wanting to guard against future unexpected uncertainties in prediction such as the increased rainfall intensity mentioned earlier).

Response: There can be a "true" parameter distribution for non-ideal problems, but only if the exact likelihood function has been used that properly considers all sources of error. Yet, such function is really difficult, perhaps impossible to derive for non-ideal cases. ABC offers several advantages here, because the summary metrics can be chosen and designed in such a way that they are insensitive to rainfall errors. Actually, we have a forthcoming paper on this topic. What we are left with in ABC is the treatment/diagnosis of model structural error. This should be easier if the behavioral ensemble is derived from summary metrics that are insensitive to rainfall errors. If one assumes that rainfall errors are unpredictable, then any pattern in the model-data mismatch must be in large part due to structural deficiencies in the model.

4764/18 more generic ABC approach – No again! It is surely GLUE that is more generic in its possibilities (including using a formal likelihood as an option where the modeler is prepared to make strong assumptions about the error structure – as demonstrated before ABC started to be more widely used in Romanowicz et al. 1994!!)

Keith Beven

Response: We do not deny that GLUE is flexible, yet many years before GLUE was proposed scientists/engineers in other fields of study were doing model-data synthesis analysis. GLUE is certainly not the first procedure that confronts models with data. We took as starting point the likelihood-free approach of Diggle and Gratton (1984), and as such GLUE is a special variant of this type of approach.

In summary, we greatly appreciate the comments of the reviewer, and will use those to our advantage when preparing our revised manuscript for publication in HESS.