Response to Reviewer 1 comments

Westoby *et al.* '*Numerical modelling of Glacial Lake Outburst Floods using physically based dam-breach models'*

Anonymous Referee #1 comments

General comment: This paper presents the results of routine work on the breach of a dam using a shallow water flow model combined with some rules for erosion estimated with the bed shear stress derived from the model. The paper is poorly written and the descriptions of the statistical methods are not rigorous and sometimes wrong. The logic surrounding a concept defined as 'equifinality' is inconsistent with the entire analysis. Given these deficiencies, the paper should not be published. Some specific details are as listed below, though this is not complete. Not all errors are listed below.
We thank the reviewer for their synopsis of the manuscript and for their specific comments. At the outset, we counter that this work is far from 'routine', particularly in the context of glacial hazard assessment, which we note that the reviewer makes virtually no specific reference to throughout their review.

The research presented remains one a handful of studies that use fully physically based numerical dam-breach models to investigate the moraine breaching problem – we view this as extremely significant, given that, despite the emergence of these models over a decade ago, the vast majority of the published literature, including many contemporary studies, still employ simple empirical, or at best, analytical models to simulate moraine dam breaching with associated, and often understated, implications for the robustness of model output (see, e.g. Westoby *et al.*, 2014).

Regarding the reviewer's stance on our definition of equifinality and its suitability for our method of analysis, we note that they have not specifically referenced the Generalised Likelihood Uncertainty Estimation (GLUE) method that we utilise, or the wealth of literature that is available to support its use (much of which is included in the manuscript). On the face of it, this could be taken to demonstrate unfamiliarity with the approach, which is a well-established method for embracing and propagating uncertainty through a modelling chain in the hydrological sciences.

We also note that the vast majority of the reviewer's comments are concerned with the dam-breach modelling component of the work. No reference is made to the photogrammetric reconstruction of the moraine topography, and a single comment addresses the two-dimensional hydrodynamic modelling component, which comprises at least a third of the research that is presented. This is at odds with the lengths to which the reviewer has gone to refute or highlight what they believe to be serious inadequacies in our approach to dam-breach modelling, and, although they state that 'Not all errors' are listed, we have taken this to imply that they have little or no objection to the those elements that they have not covered in their review.

We address their specific comments below.

Specific comments:

1) The concept of equifinality, used on line 19, page 481 before it is defined on page 482, is defined by the authors as "identical or similar end states in an open system are achieved from many different combinations of input parameters, initial conditions, and model structures" (lines 1-3, page 482). The logic here is inconsistent with the authors' later application of maximum likelihood (section 4.3, page 486), since maximum likelihood is

based on the concept that differences between input parameters in a model produce differences in the output of the model. If they do not, then the parameters have no relevance to the data being modelled. As defined, the existence of 'equifinality' means that, if you are varying the input parameters over a large enough range then you still get the same output. Physically, this means that you have the wrong model. With equifinality, there is no basis for statistical inference or maximum likelihood.
The reviewer's comment regarding the contradiction between the concepts of equifinality and maximum likelihood is correct, and the error here lies in our use of the term 'maximum likelihood', as opposed to simply 'likelihood'. In the original description of the GLUE method (Beven and Binley, 1992, p281), the term 'likelihood', is used in an extremely general sense as a 'fuzzy, belief, or possibilistic measure of how well the model conforms to the observed behaviour of the system, and not in the restricted sense of maximum likelihood theory…'
The fault that the reviewer points out lies in our description of the method used to quantify model, or parameter ensemble, performance. It is not maximum likelihood theory that we apply (which, as the reviewer correctly points out) would be incompatible with equifinal output), and we maintain that our approach of assigning positive likelihood values to behavioural model runs is entirely consistent with the GLUE approach and is methodologically sound. Our terminology will be corrected in the revised manuscript and the differences between maximum likelihood and the method of quantifying model performance and assigning likelihoods will be made absolutely clear.

2) Lines 3-4, page 483. Is this paper about equifinality or the null space for this model? Or is it about a dam breach study?
The paper seeks to demonstrate how numerical models representing difference phases, or components, of the GLOF process cascade can be chained together to provide probabilistic predictions of the breaching process and downstream flood propagation. We use established models for this purpose, but employ Generalised Likelihood Uncertainty Estimation to quantify the predictive uncertainty associated with this entire modelling process.

3) Section 4.2: There are no details of what HR BREACH does or how it works. Why are journal subscribers paying to advertise a proprietary model that is marketed for personal gain? There is no way to assess the efficacy of this model with the data presented here, since the data are from a field example of a breached moraine dam. The data from a lab experiment that was done is mentioned, and these results do not support the model as I discuss below. What is the benefit to the scientific community of a proprietary model advertisement? Why not just buy an ad for the journal?
It is our opinion that enough detail is included so that the reader is able to understand how the model is set up, executed, and the various process equations that it uses in the breaching calculation. As stated in the manuscript (p9, line 7 onwards), HR BREACH employs flow hydraulics, sediment erosion and discrete embankment stability analysis to simulate breach development, and has not been directly calibrated to datasets from field- or laboratory experiments. This makes it particularly unique in the context of its appropriateness for moraine dam breaching simulation, and sets it apart from the vast majority of published studies which have used simple, empirically-derived models and fail to discuss the (sometimes highly questionable) suitability of their application. The revised manuscript will make our decision to use HR BREACH clearer in this respect, and will contain an expanded description of precisely how the model simulates breach development based on the various equations that are already presented.

We believe that the remainder of this comment has no place in the review, and serves little purpose other than to highlight the reviewer's apparent wholesale opposition to the use of proprietary modelling software by the scientific community. This paper is NOT intended to be an advertisement for HR BREACH, and we do not apologise for its use. To suggest that its use and documentation of its capability in the context of moraine dam breaching serves 'no benefit to the scientific community' is counterproductive and off-target, and we would hope *that the Associate Editor and many readers would agree on this point. We note that the reviewer has made a point to single out HR BREACH here, and makes no reference to the appropriateness of our use of, for example, ISIS 2D for hydrodynamic modelling, or ArcGIS for generating topographic grid topography, both of which are also proprietary, and to which they could have equally referred.*

To be absolutely clear, our desire here is to highlight the value of such models for moraine dam breaching simulation – at the time that the research was conducted, HR BREACH represented virtually the only physically based dam-breach model that was available to conduct the research with the resources available and within the time-frame available, and represented a significant improvement in terms of model physicality over existing models.

4) Model dimensionality is mentioned, and this is an issue here. There are 4 free parameters in equation 1, at least 7 in equation 2 (assuming the weights are measured), an additional 2 in equation 3, and there are at least 4 parameters (including bed friction) in the shallow water flow vector in the code used to analyze the dam breach. This adds up to seventeen free parameters to constrain with the dam breach data, which here consists solely of the initial dam height and depth, and the change in the thalweg elevation of the breach with time. A standard Bayesian analysis would show that the evidence from these thalweg and height measurements are spread thinly over the model space – in other words that the data constraints on their model are weak.
We agree with the reviewer here that the model chain is parametrically complex. However, by using methods to explore the parameter space and compare the emergent responses to identifiable observables (i.e. our geometric, evaluative measures of model performance), we offer the first field-based assessment of the application of numerical models to moraine breaching and outburst flooding. The imbalance between the degrees of freedom in the model parameter space and objective test data is common to many areas of numerical modelling in the Earth surface sciences (e.g. rainfall-runoff and flood inundation modelling), and GLUE provides a useful method to explore the resulting uncertainty that arises from such ill-conditioning. Moreover, our approach provides tools to present these results openly and objectively as shown with respect to the parameter values in Fig. 7 and the range of possible model performance shown in Figs. 8, 9, 10 and 13.

5) Discussion of Monte Carlo sampling and clustering, pg 487-488: Use of Latin hypercube sampling would accelerate the search and eliminate the 'clustering'. This discussion is confused, since without additional criteria, all random walks will produce a 'cluster' of samples in the area where the search is started before wandering away from the starting point. The number of samples required to be representative depends upon the problem being studied and where the search is started, so the number reported here is irrelevant to other dam breach studies.
We thank the reviewer for highlighting the effectiveness of Latin Hypercube sampling for eliminating clustering. Monte Carlo sampling capabilities are in-built in HR BREACH, with no capability to use other stochastic sampling methods, hence the reason for its use here. It is also a well-established sampling approach that is wholly compatible with the GLUE method, as stated in the paper.

Yes, the number of model runs required for a representative result will vary depending upon the specific breaching problem under investigation, but at no point in the text do we imply that the one thousand runs that we undertook would be directly applicable to other situations. To avoid any confusion, we will re-word the final sentence in this paragraph, which currently reads 'This number of simulations was therefore deemed to be acceptable for stochastic sampling' to make it clear that this applies to this study only, and that the number required will likely vary for other studies. We will also add a discussion of different stochastic sampling methods, including Latin Hypercube sampling, to widen the scope of this paragraph and the approach in general.

6) Lines 23-28, page 488: "ensembles were assigned positive likelihood values in the range 0-1". Assigned? Likelihood is calculated by comparing the model output to model input data, which if equifinality applied would give you nothing. And for Bayesian inference, which the authors' later claim on line 11 page 490, the likelihood is modified by prior information and normalized by the integral over all possible posterior distributions of input parameters, sometimes called the evidence integral. Bayes equation and standard Bayes updates Mackay, 2003, Information Theory, Inference, and Learning Algorithms, Cambridge Univ. Press) bears no resemblance to what is presented here.
We refer here to our response to point (5) and our inappropriate use of the term 'maximum likelihood' and refer the reviewer to the various GLUE literature that is referenced in the manuscript.

7) Equation 4, page 489: With equifinality, there would be no relevance to this equation, or reason to use it.
The reviewer seems to have taken an incredibly literal stance on the definition of equifinality, assuming it to describe <u>identical</u> system end states, whereas the GLUE method, and our introduction to it on p482, describes how it may equally refer to <u>similar</u> end states. In line with GLUE theory, it is the latter that is considered in this study. Perhaps we have not made this clear at the outset, and so the revised manuscript will make this definition and our treatment of the equifinality thesis absolutely clear.

The equation that the reviewer refers to is used as a measure for evaluating modelled thalweg elevation profile, as compared with that observed in the field. We maintain that its use remains appropriate in this context, as it is used to assign the highest likelihood scores to those runs that perform the best at reproducing the observed thalweg topography, whereas those that perform the poorest are assigned the lowest values – this is consistent with our two other methods of evaluating ensemble performance.

8) Line 23, page 489: What linear likelihood function? Not equation 4!

We will add a description of the linear likelihood function to the revised manuscript.

9) Lines 24-28, page 489: A dam breach is a three dimensional process, as demonstrated in all dam breach experiments. What is the justification of the two dimensional approximation? How is this useful to the reader?
Yes, the breaching process is a fully three-dimensional problem. However, the current state of the art in breach modelling employs 2D approximations; in part as a reflection of available computing resources, but also, and arguably more profoundly, as a reflection on the dearth of data that are available for their parameterization. We note that most applications of breach models are designed for linear embankments, where spatial flow concentration is negligible. In our case, the curved nature of the moraine may play an important role in the response of the natural system. The fundamental difference between the intended applications of these models (i.e. man-made constructions) and our case study (a natural, highly irregular feature) is a key point, and warrants further discussion which will be provided in the revised manuscript. The 2D metric that we use, and which we believe the reviewer is referring to, relates less to the lateral position on the moraine where the breach emerges, as to the longitudinal growth of the breach through the moraine. It is our unique 3D SfM topographic model of the moraine that enables us to provide quantitative data with which to test these predictions – this approach is unlike any comparable existing field scale research undertaken in this field to date, and is a novel and important contribution of the paper.

As a side note – the reviewer could equally have questioned our use of a 2D approximation (ISIS 2D) for simulating downstream flood propagation, which we note that they have not raised. A similar response would apply here – to expect nothing less than the use of a 3D *hydrodynamic solver would be unrealistic and entirely impractical given, amongst other things, the required computational burden and present software availability and cost.*

10) Equations 5 and 6: For a Bayesian update, equation 6 should be an integral, as should the numerator quantities in equation 5. As written these equations are undefined and useless to the reader.

We will clarify the mathematical presentation of the methods used.

11) Section 4.3.4, page 491: Again, this likelihood construction is inconsistent with the authors' overriding concept of 'equifinality'.
In our opinion, this comment serves to further highlight the reviewer's apparent lack of understanding or familiarity with the fundamentals of the GLUE method, wherein the construction of cumulative distribution functions is required for the subsequent analysis and extraction of prediction percentiles, or confidence limits. GLUE theory, and examples of its use in a variety of geoscience sub-disciplines, is supported by a vast number of peer-reviewed publications, the majority of which employ the basic workflow that we describe in this manuscript. We would suggest that the reviewer refer to this literature, the most relevant of which (e.g. Beven and Binley, 1992, Beven and Freer, 2001, or Beven, 2005) are referenced at numerous points throughout our methods sections, and none of which the reviewer has made reference to in their review.

12) Line 19, page 491: "the modeled spillway measured 0.5m wide and 0.5 m deep". The results from this experiment, introduced and briefly described here, are summarized on page 496, after interleaving a discussion of the field example modelling. This field discussion is a non-sequitur and very confusing. As for the comparison with the lab experiment, the model produces discharges an order of magnitude larger than were observed for the same boundary value problem. This is very poor model performance, and raises questions about the viability of the field simulation. I fully expect that there would be good agreement with the inundation in a field comparison if there were good surface control, since this depends primarily upon continuity. However the timing of any simulation would be very different from what actually happened, making the application of this model useless for hazards estimation.
Our solution to the first point raised by the reviewer here would be to add another sub-heading, which would become 4.4.1, and would therefore be consistent with our approach of describing the two types of 'field example' modelling within their own respective sub-sections. It is unclear as to whether the reviewer is implying that the purpose of the field example (i.e. overtopping and mass failure) modelling is unclear, or our description of each. We suspect that it is the former, and acknowledge that this section is lacking a clear introductory paragraph that sets out our reasoning for introducing the model perturbations. This will be addressed in the revised manuscript.

Regarding the latter part of this comment, which focuses on model performance and modelled discharge magnitudes, it is only the control and overtopping scenarios that share the same initial boundary values or conditions. The fundamental difference here is the interaction of the overtopping waves with the dam structure in the earliest stages of the simulation, which initiate rapid down-cutting of the moraine and facilitate the removal of larger quantities of water. In contrast, the instantaneous mass removal scenarios do not share the same boundary conditions with the control experiment, since the modelled spillway dimensions are far higher (with cross-sections of 1, 3, and 5 m^2 along the spillway length). In a similar, but more extreme, version of what occurs with the addition of overtopping waves, the lake pressure head therefore produces far higher peak discharges for the overtopping scenarios, as less energy is required for initial breach down-cutting and expansion. Having said this, the viability and realism of this particular field experiment remains the weakest of the three modelling scenarios since HR BREACH is incapable of simulating the specific mass failure mechanisms associated with this type of moraine breaching, hence our *approximation by modifying the initial spillway dimensions. More could be made of the significance of this, and it will be discussed in detail in the revised manuscript.*

In this analysis described here, the word 'behavioral' means 'realistic'. Results deemed not 'realistic' were thrown out. Why? This is the antithesis of Bayesian inference, and to even suggest that this study is a Bayesian analysis is grossly misleading. The truly valuable result is that only Manning's '*n*' is significant. Given how the authors have set up their problem, with equation 1 governing erosion, this result is expected.
Once more, we refer the reviewer to the basic tenets of the GLUE method and supporting literature, wherein the relative likelihood of a given model and parameter set in reproducing datasets available for model evaluation (in this case, geometric descriptors of dam morphology) is achieved by effectively 'rejecting' those parameter sets that are deemed unacceptable or non-behavioural by assigning them a likelihood score of zero. We employ Bayes Theorem here to conjoin, or update the likelihood score of each retained parameter ensemble. Using this method, an existing probability is updated using additional information to calculate a conditional probably that incorporates both measures, resulting in a single likelihood score that reflects multiple, separate weighted likelihood scores.

Yes, Manning's n *appears to govern erosion and is the key parameter here, but we argue that it is not only the truly valuable result. Considering its significance for the reconstruction of moraine breaching, and specifically the various triggering mechanisms that could initiate dam failure, the study has demonstrated that near-identical breach morphologies can be produced by a range of breach initiation mechanisms, including overtopping waves of varying magnitude. This is significant for dam-breach reconstructions, which typically use field-based observations such as high-water marks and strandlines in an effort to constrain peak discharge, with little or no consideration of the timing of flood passage (assuming the absence of first-hand or instrumented observations of the outburst). One of the novel contributions of this research is the use of morphological descriptors of the breached moraine dam to condition the results of a numerical dam-breach model for outburst flood reconstruction, which is of particular value in situations where field measurement of outburst flood deposits on the floodplain is dangerous or simply impossible due to logistical considerations.*

13) page 495: The severe oscillatory behavior the authors observed in the ADI method could be due to carbuncles cropping up in the solution, when a shock is aligned with the grid. An ADI method would amplify these.
We thank the reviewer for pointing this out, and will include this as a possible explanation for the observed flow depth oscillations that were observed.

14) line 1, page 501: Maximum likelihood is effectively identical to soft K means clustering (Mackay, 2003, Information theory, Inference, and Learning Algorithms, Cambridge Univ Press, chapter 22), and should be referenced and discussed here.
We will update the manuscript appropriately and include supporting references as suggested.

15) Line 17-22, page 504: Manning's 'n' coefficient of .02 - .029 corresponds to an effective median grain size of between 0.03 m and 0.2 m (Henderson, 1966, Open Channel Flow), which fits with the authors' bed observations. These values are not low (as the authors' state), except in the application of poor models where high values of bed resistance are needed to stabilize the model calculations. And, as the authors state, the Dig Tsho moraine is composed of gravel, cobble, and boulder-sized material, whose median grain sizes do fit their model result. This is the only significant result presented in this paper.
The reviewer's stance here on behavioural Manning's n *coefficients appears to imply that, if additional resistance effects (i.e. form, spill, curvature) are fully accounted for, the resistance term should reduce to account for the particle (or 'skin') resistance alone. This is correct, and*

would result in what we have described as 'low' resistance values. However, since 2D fluid flow is modelled in HR BREACH at comparatively coarse grid resolutions, it is reasonable to expect that the 'effective' values of n will incorporate these macroscopic properties. Our results suggest that a reduction in n *is more likely to reflect parameter compensation (or interaction) effects elsewhere in the model chain – e.g. a reduction in* n *will increase shear stress, which could offset other parameters that reduce bed erodibility. Given the complexity of the model parameter space, we maintain that it is these effects that account for the low values of* n*, rather than uniquely the identification of appropriately scaled particle roughness values.*

Supporting references

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

Bevan K and Freer J (2001) Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. *Journal of Hydrology* 249, 11-29.

Beven K (2005) A manifesto for the equifinality thesis. *Journal of Hydrology* 320, 18-36.

Westoby MJ *et al.* (2014) Modelling outburst floods from moraine-dammed glacial lakes. *Earth-Science Reviews* 134, 137-159.