Referee 1: Tobi Krueger

This is a short technical note on the first release of an uncertainty estimation toolbox for Matlab. The toolbox itself has very useful features, most notably from my perspective: advanced multivariate distributional assumptions through copulas; an audit trail log of assumptions made in the process; and 12 workflow scripts that will help the analyst apply the toolbox.

In my opinion it would increase the value of the paper (as an addition to the documentation of the toolbox) if the authors could give a little more detail on the nature of the uncertainty problem in each workflow. Now, if the paper is the first port of call for those wanting to use the toolbox, they have to do a lot of additional background reading before they can decide which method to use or which workflow to follow. If the authors could give a little more detail on the nature of uncertainties considered in each workflow and why this led them to use a certain method and not another. This would then give the reader a sense of which method is appropriate for what.

The organisation of the paper has been revised to give more prominence to the discussion of methods and choice of workflow (former Figure 6) earlier in the paper.

The description of the toolbox is otherwise clear, with the exception of the handling of input data uncertainty in the “conditioned” case. See also my comment to Fig6 below: Is input data uncertainty handled the same way as parameter priors? There should be added complexity due to the timeseries nature of inputs. What are the choices available to the analyst here? I’m thinking of rainfall multipliers and various other approaches that have been suggested. And are these prior choices of input data uncertainty updated conditional on the observations of model output? I think here we need more information in the paper.

The text and Table 1 have now been revised to make the choices clearer.

Specific comments:

There is a font size change from L75 onwards. Is this intentional?

No, this has now been changed

When reviewing the other toolboxes from L75 onwards, could the authors flag those toolboxes that are still maintained? With some I had the feeling they might not be (certainly DUE) and then they are of limited value in my opinion.

That is a good point. We have added a comment that this is the case, though it is not always possible to tell from the sites as to whether they are still maintained or not.

On the CURE website there is still what seems like an older version of the paper submitted to EMS. This should be updated.

This will be updated when the revised paper is submitted

Fig3-4: The font sizes and tick labels (overlap) could be improved in some instances.

This has been corrected

Fig3: I find the evolution plot of the Rhat statistic not so useful since we can't see the area around 1 very well at that scale where we want Rhat to end up. I encourage the authors to consider removing this plot or finding a way to scale it better (I guess the final Rhat value is reported in any case, which would be important).

These was a result of the way in which the figure was displayed. It is now improved.

Fig6: I'm unclear about the two branches coming from "Conditioned Uncertainty Analysis". Is the analyst meant to go down both branches? If so, this could be made clear. This way the analyst seems to end up with prior parameter distributions at the end of the left branch that then feed into the sampling strategies emerging from the right branch. But what about the multivariate cases – what do they lead to? The same could be said about the middle arrow on the very left (forward uncertainty analysis). More importantly, what about input data uncertainty? Is this handled the same way as parameter priors? There should be added complexity due to the timeseries nature of inputs. What are the choices available to the analyst here? I'm thinking of rainfall multipliers and various other approaches that have been suggested. And are these prior choices of input data uncertainty updated conditional on the observations of model output?

This is certainly ambiguous and has now been clarified in the text and in a revised Figure 6 which now makes those two branches clearer.

Comments for general discussion:

I'm here noting a few general discussion points that I invite that the authors to engage with, though they are not of central importance to the paper.

First, I want to encourage the authors to eventually publish their toolbox for an open-source software environment like R as well. Or at least comment on the compatibility of the toolbox with clones like GNU Octave.

We would certainly like to do so (as with the earlier highly successful SAFE Toolbox, Pianosi et al EMS 2016), given time. If it is possible to add other versions then they will be made available as open source on the CURE site.

Pianosi, F., J. Rougier, J. Freer, J. Hall, D. B. Stephenson, K. J. Beven, and T. Wagener, 2016, Sensitivity Analysis of environmental models: a systematic review with practical workflows, *Environmental Modelling and Software,* 79: 214-232

Second, I'm increasingly wondering whether the distinction between formal and informal uncertainty methods is really productive (I say this having used this distinction myself). This terminology once served to circumvent accusations of incoherence of methods like GLUE just by introducing a different label (informal), but now distracts from engaging with what really matters: the assumptions made about various sources of uncertainty when using certain methods (a problem that the authors summarise well by the way and tackle through their audit trail). With the formal/informal terminology one is led to believe one has a choice between formal and informal methods, while in reality in both cases one has a much more difficult choice of how exactly to model uncertainties and how to aggregate individual (e.g. timestep-based) performance measures into an overall metric (e.g. multiplicative or additive) and what understanding of uncertainties and model performance this entails (often implicitly). Such understanding goes down to foundational axioms as discussed by Nearing et al. (2016), which the authors cite. Any

formal/informal distinction would also be increasingly blurred by methods such as ABC – here I'm missing reference to work by Lucy Marshall and co-workers discussing the similarities between ABC and GLUE, by the way. Maybe the authors can re-evaluate their use of these terms and engage in a broader discussion.

This is an important point, but, we would suggest, not as clear-cut as Nearing et al suggest. Yes, if you are happy to fit into the axioms of probability theory then there is a fundamental foundation to all that follows. But that assumes that the probabilities are complete and non-conflicting (the excluded middle). One way round this is to allow alternative probabilistic representations in parallel (e.g. Rougier and Beven, 2013) but there are alternative sets of axioms that could be considered, such as those underlying fuzzy set theory (e.g. the Halpen 2003 book). The point we wish to make is that there are alternative assumptions (within broadly being Bayesian in conditioning the outputs by information) but that the assumptions made must be made explicit since, for the case of epistemic uncertainties, there can be no right answer (Beven et al., NHESS 2018a,b). We would suggest that Bayes' use of subjective odds in his original paper, is compatible with this approach without the need for formal likelihood reasoning.

Rougier, J and Beven, K J, 2013, Model limitations: the sources and implications of epistemic uncertainty, in Rougier J, Sparks, S and Hill, L (Eds), *Risk and uncertainty assessment for natural hazards*, Cambridge University Press: Cambridge, UK, 40-63

Beven, K. J., Almeida, S., Aspinall, W. P., Bates, P. D., Blazkova, S., Borgomeo, E., Freer, J., Goda, K., Hall, J. W., Phillips, J. C., Simpson, M., Smith, P. J., Stephenson, D. B., Wagener, T., Watson, M., and Wilkins, K. L., 2018, Epistemic uncertainties and natural hazard risk assessment. 1. A review of different natural hazard areas, *Natural Hazards and Earth System Science,* 18(10): 2741-2768. https://doi.org/10.5194/nhess-18-2741-2018

Beven, K J, Aspinall, W P, Bates, P D, Borgomeo, E, Goda, K, Hall, J W, Page, T, Phillips, J C, Simpson, M, Smith, P J, Wagener, T and Watson, M, 2018, Epistemic uncertainties and natural hazard risk assessment – Part 2: What should constitute good practice?, *Natural Hazards and Earth System Science*, , 18(10): 2769-2783, https://doi.org/10.5194/nhess-18-2769-2018

Third, I find the comment about rigorous statistical treatment of aleatory but not epistemic uncertainty in L59-61 misleading. This seems to be relating more to differences between frequentist and Bayesian statistical methods than anything else. The Bayesian framework deals expressly with epistemic uncertainties. The question is just whether or not the assumptions one makes are justified – but this is the case with any uncertainty method (which the authors emphasise well in this paper). I encourage the authors to remove this reference to epistemic versus aleatory uncertainty and focus on the importance of choices (which they already do) – in any method.

Again we are emphasising that importance of making assumptions explicit. But there is a difference. Under aleatory assumptions you can develop a formal likelihood function for use within Bayes equation. But we know that for many cases of practical interest where epistemic uncertainties are important this gives misleading results by stretching the likelihood surface (in some cases of time series enormously). Although those favouring the probabilistic approach might suggest that it is then the definition of the likelihood that needs improving (e.g. the recent paper of Vrugt et al., JH2022 suggesting a "universal" likelihood that still ignores sources of epistemic uncertainty except in so far as they affect the statistics of the residuals). But, we would suggest that it might be possible to be more thoughtful than that, for example in trying to use knowledge about the observations to define limits of acceptability, eg. The recent paper by Beven et al. (HP2022).

Vrugt, J.A., de Oliveira, D.Y., Schoups, G. and Diks, C.G., 2022. On the use of distribution-adaptive likelihood functions: Generalized and universal likelihood functions, scoring rules and multi-criteria ranking. *Journal of Hydrology*, *615*, p.128542.

Beven, K. J., Lane, S., Page, T., Hankin, B, Kretzschmar, A., Smith, P. J., Chappell, N., 2022, On (in)validating environmental models. 2. Implementation of the Turing-like Test to modelling hydrological processes, *Hydrological Processes*, 36(10), e14703, **https://doi.org/10.1002/hyp.14703**.