Review of the Technical Note "The CREDIBLE Uncertainty Estimation (CURE) toolbox: facilitating the communication of epistemic uncertainty" by Page et al.

The technical note by Page et al. presents an open-source MATLAB toolbox that combines multiple uncertainty estimation methods. Workflows are provided that exemplify the application of each of these methods. The toolbox also allows the use of a condition tree that enables users to document their choices and assumptions in an attempt to improve transparency and communication of the limitations of the results. While I recognize the value of such a toolbox and appreciate the effort in compiling all these uncertainty estimation methods, I believe a thorough revision of the code and significant improvements to the documentation need to be made. More detailed documentation has the potential to increase the use of the toolbox substantially and may prevent its use in unintended ways (with misleading results/conclusions). I have listed the main thoughts about the paper/code below.

1. Evaluation against the original implementation

Given some of the results I saw (for example, the results from Workflow 6), I wonder whether each uncertainty estimation method was tested against their original implementation (i.e., by running the CURE toolbox and the original code using the exact same case study). The authors could choose as case studies for the workflows the examples provided with the original (cited) toolbox (whenever possible), which would allow the user to know the expected results and help interpret the outputs provided by the CURE toolbox.

2. CURE toolbox vs. original implementation

I don't believe the codes reflect the implementation described in the references cited in Table 1. For some uncertainty estimation methods, the current implementation is a simplified version of the original code with limited options. This should be explicitly mentioned in the manuscript, and a discussion of potential implications should be incorporated.

3. Large amount of typos in the codes

Codes and website content need to be revised carefully. There are multiple typos in the codes that, in some cases, may prevent the user from running a specified configuration (for example, I get an error when running Workflow 6 by setting “Err_mod_fit = 1” and switching the option “Lifun” from “B_C_AR” to “B_C”).

4. Need to improve the documentation

I am not sure if the intended audience would be able to use the methods implemented in the toolbox without substantial additional effort.

It is not clear what are the requirements of each uncertainty estimation method (which inputs are needed, what are the user-specified options) or, in other words, which parts of the code the user should modify when creating their own workflow.

A similar comment is valid for the outputs: some output figures do not have appropriate titles/legends; it is not always clear what is the meaning of each figure; and why these figures are being plotted.

It would be helpful to include in the manuscript a table (or something similar) where the required inputs, available options (and what happens when the user selects each option), and outputs of each uncertainty method are at least listed (even better if explanations of the generated outputs are provided).

5. Treatment of autocorrelation

I am not sure if the autocorrelation implementation is correct. Any references? I would suggest checking Evin et al. (2013, 2014) and Vrugt et al. (2022).

> G. Evin, D. Kavetski, M. Thyer, and G. Kuczera. Pitfalls and improvements in the joint inference of heteroscedasticity and autocorrelation in hydrological model calibration. Water Resources Research, 49(7):4518–4524, 2013.
>
> G. Evin, M. Thyer, D. Kavetski, D. McInerney, and G. Kuczera. Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity. Water Resources Research, 50(3):2350–2375, 2014.
>
> Vrugt, J. A., de Oliveira, D. Y., Schoups, G. & Diks, C. G. On the use of distribution-adaptive likelihood functions: Generalized and universal likelihood functions, scoring rules and multi-criteria ranking. Journal of Hydrology 615, 128542. doi:10.1016/j.jhydrol.2022.128542 (2022).

6. Limitations/advantages of each method and additional guidelines

Some guidance could be provided for selecting from the different uncertainty estimation methods available. A summary of the limitations/advantages of each method would be helpful, as well as guidelines for the specification of the required parameters (for example, studies like McInerney et al., 2017 could be provided as a reference for helping users to select initial Box-Cox parameter values).

> McInerney, D., Thyer, M., Kavetski, D., Lerat, J., and Kuczera, G. (2017), Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors, Water Resour. Res., 53, 2199– 2239, doi:10.1002/2016WR019168.

7. Going beyond recording assumptions

The toolbox includes a condition tree that users can use to document choices and assumptions. It would beneficial if the toolbox also included the option of testing (some of) these assumptions. For example, if the uncertainty analysis is conducted using a formal likelihood/Bayesian inference, common checks would be to evaluate the residual assumptions and the quality of the uncertainty estimates (see Thyer et al., 2009 and many others).

> M. Thyer, B. Renard, D. Kavetski, G. Kuczera, S. W. Franks, and S. Srikanthan. Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis. Water Resources Research, 45(12), 2009.

While this might be beyond the scope of the paper, at least some discussion on how the assumptions can be tested could be included, especially for "informal" metrics (given that in some cases the assumptions are not known, how can we evaluate the reliability of the results?).

Minor comments

1. What is the difference between the GLUE-Limits of Acceptability (LoA) method and the use of LoA within the DREAM toolbox (Vrugt & Beven, 2018)?

   > J. A. Vrugt and K. J. Beven. Embracing equifinality with efficiency: Limits of acceptability sampling using the DREAM(LOA) algorithm. Journal of Hydrology, 559:954–971, 2018

2. McInerney et al. (2018) could be added to the introduction.

   D. McInerney, M. Thyer, D. Kavetski, B. Bennett, J. Lerat, M. Gibbs, and G. Kuczera. A simplified approach to produce probabilistic hydrological model predictions. Environmental Modelling Software, 109:306–314, 2018.

3. Residuals and errors are used interchangeably, but they do not have the same meaning. From Vrugt & de Oliveira (2022): “The word error implies a difference between an observed value and its true value. As our measurements of system behavior are imperfect, the residuals are estimates of the errors under the assumed model. Hence, we should use the word residual instead.”

   Vrugt, J. A. & de Oliveira, D. Y. Confidence intervals of the Kling-Gupta efficiency. Journal of Hydrology 612, 127968. doi:10.1016/j.jhydrol.2022.127968 (2022).

4. Glossary: some definitions are somewhat informal and not necessarily accurate. Even though the toolbox was designed for individuals “who are not necessarily experts in uncertainty estimation”, it is important to define the terms accurately.
5. Code, Workflow 1: “This Workflow is a CASE STUDY demonstrating forward uncertainty analysis for the CHASM landslide model (see CHASM_IO_files_2014.pdf in CHASM folder)”: I couldn’t find the “CHASM_IO_files_2014.pdf” file
6. What is the difference between a “case study” (workflow) and “application example”/”example” (other workflows)?
7. Website, Learn about CURE and Uncertainty Estimation: Update link and reference to Page et al. (2021)
8. Website, Case Studies: Why only 4, 5, and 11 are presented here?