Thank you very much for doing this work. I think the paper could make a contribution to HESS after some work. I leave here my main comments, I hope they help to improve the manuscript.

**Response:**

Thank you for your valuable feedback. We appreciate your comments and will address each suggestion with corresponding revisions to improve the manuscript.

**Comment #1:** Literature review. I miss references to work of Jasper Vrugt, mainly to Vrugt et al., 2012 (the DREAM paper), you might have a look at the work of Carlo Alber about ABC(during 2015 and maybe reply rom Vrugt later on), Kavetski et al. 2018 and Fenizia et al 2018 look through the introduction+ references of these 2 papers;and the work from Prieto et al. 2021;2022 about hydrological mechanisms identification and the diagnostic metrics there in. Othersize the introduction is a bit repetitive and convoluted but it misses inormation, e.g. about the choose or development of different likelihoods (later on you assume a normal gaussian but I do not see the justification), model diagnostics metrics and why.

**Response:**

We appreciate your suggestion to reference Jasper Vrugt et al's 2012 work on the DREAM algorithm, as well as Carlo Albert's research on the ABC method, particularly his 2015 study and subsequent discussions. We also acknowledge the contributions of Kavetski et al. and Fenizia et al. in 2018. These studies are highly relevant to our topic, and we will review and incorporate them into the revised manuscript to strengthen the comprehensiveness of our literature review.

Regarding the likelihood function, your feedback is valuable. We currently use a Gaussian distribution but have not sufficiently explained its theoretical basis. In the revised manuscript, we will clarify the rationale for this choice, and its relevance to our algorithm, and provide a comparison with other commonly used likelihood functions to explain our final selection.

2. benchmark: I am missing a benchmark to compare. Maybe, one good idea might be to use the package from Vrugt for DREAM as benchmark – I am aware the author had everything ready to be applied.

**Response:**

Regarding the benchmark testing of the PEM-SMC algorithm, we evaluated its effectiveness through two experiments, as presented in Supplementary Information S1, "Evaluating the effectiveness and efficiency of the revised PEM-SMC algorithm". However, we recognize that these internal tests primarily demonstrate the algorithm's capability to estimate the target posterior distribution. A more comprehensive evaluation of its performance, including residual characteristics, confidence intervals, and computational speed, requires comparison with other algorithms, which is currently missing. Therefore, in the revised manuscript, we will conduct a comparative analysis between the PEM-SMC and DREAM algorithms to provide a more thorough evaluation of PEM-SMC's performance. Additionally, we will also reassess the structure and framework of the PEM-SMC algorithms in relation to the DREAM algorithm.

3. For the equations I suggest to use a properly maths notation. At least for me is helpful and I am sure for readers too. E.g. why is everything italic? I suggest you to differenciate vectors, matrices, random variables, etc (ie bold, capital letters, and so on). In Prieto et al., 2019; 2021; 2022 you can find examples for this and in whatever paper from Vrugt I am sure too.

**Response：**

Thank you for your suggestion. We acknowledge the current manuscript's limitations in the representation of equation symbols and variables, particularly the use of italics for all elements. In response, we will revise the manuscript to adopt a more standardized mathematical notation. Specifically, we will use bold italics to distinguish vectors and matrices, and uppercase letters for random variables, to enhance clarity and readability. We will refer to the works of Prieto et al. (2019, 2021, 2022) and Vrugt's research to ensure consistent and standardized symbol usage for the benefit of the readers.

4. Posterior pdf: please for the posterior of the parametere use the full Bayes equation and then say that the left hand is proportional to the right hand so that non Bayesian can follow it. Indeed this is meet bacause you use only one model (eg see prieto et al., 2021, 2022).

**Response:**

Thank you for your suggestion. Following the work of Prieto et al. (2021), we have revised the posterior probability density function (PDF) of the parameters in the original manuscript and provided a more detailed explanation of the full Bayesian equation. To improve clarity, we have also incorporated a conceptual model representation, further elaborating on how the likelihood function, when combined with observed data and prior knowledge, results in the derivation of the posterior parameter distribution. The revised section is as follows:

*In this study, we address the inference of model parameters in CoLM using observed data $\tilde{q} = (\tilde{q}_t; t = 1, ..., N_t)$, which represents a time series of target variable observations (e.g., LE and NEE) of length $N_t$. Within the Bayesian framework, model parameters are conceptualized as probabilistic variables, and the posterior distribution of the parameters, $p(\theta \mid \tilde{q}, G, \tilde{x}, s_0)$, is expressed as:*

$$p(\theta \mid \tilde{q}) = \frac{p(\theta)p(\tilde{q}|\theta)}{p(\tilde{q})} = \frac{p(\theta|G,\tilde{x},s_0)p(\tilde{q}|\theta,G,\tilde{x},s_0)}{p(\tilde{q}|G,\tilde{x},s_0)}$$

*Where $p(\theta)$ is the prior distribution of its parameters over its feasible domain, $p(\tilde{q}|\theta)$ is the likelihood function associated with the probability mode and $p(\tilde{q})$ is referred to as Bayesian Model Evidence (BME) or Marginal Likelihood. The BME term is generally the normalization constant and is not required for parameter inference. The model structure $G$, forcing data $\tilde{x} = (\tilde{x}_t; t = 1, ..., N_t)$, and initial conditions $s_0$ are treated as fixed in this study.*

5. – concern: why the likelihood is a normal likelihood? could you please justify and then analize the residuals of the posteriors? are you also meaning that the vairables are independent and then the likelihoods can bu multiplied? I suggest you to have a look at ABC hear just in case it can help.

**Response:**

Thank you for your valuable feedback and suggestions.

After conducting a posterior analysis, we found that the residuals of the LE and NEE variables align more closely with a t-distribution, rather than the initially assumed normal distribution (Figure 1). The normal distribution was our initial choice due to its

computational simplicity and widespread use. However, the t-distribution, with its heavier tails, is more sensitive to outliers and may better fit our current dataset. While there are notable differences between the likelihoods of the normal and t-distributions, particularly in handling extreme values, their roles in parameter optimization are quite similar. Therefore, we believe that this discrepancy is unlikely to significantly impact the optimization results. Nevertheless, to ensure the robustness of our findings, we plan to use the t-distribution likelihood function in future work and compare its performance with that of the normal distribution to validate this assumption.

We assume that the observations are independent, allowing the likelihood function to be expressed as a product of the probabilities for each independent observation. This independence assumption simplifies the computational complexity and ensures the interpretability of the model. Given the current structure and data of the model, this assumption appears reasonable. However, if dependencies between observations emerge in future models, we will revisit and adjust this assumption accordingly.

We also appreciate your suggestion regarding the use of Approximate Bayesian Computation (ABC). ABC is particularly useful in cases where the likelihood function is intractable, providing an effective approach for parameter estimation in complex models. Although we can compute the likelihood function directly in the current study, we will thoroughly review relevant literature and explore the potential application of ABC in more complex scenarios in future research.
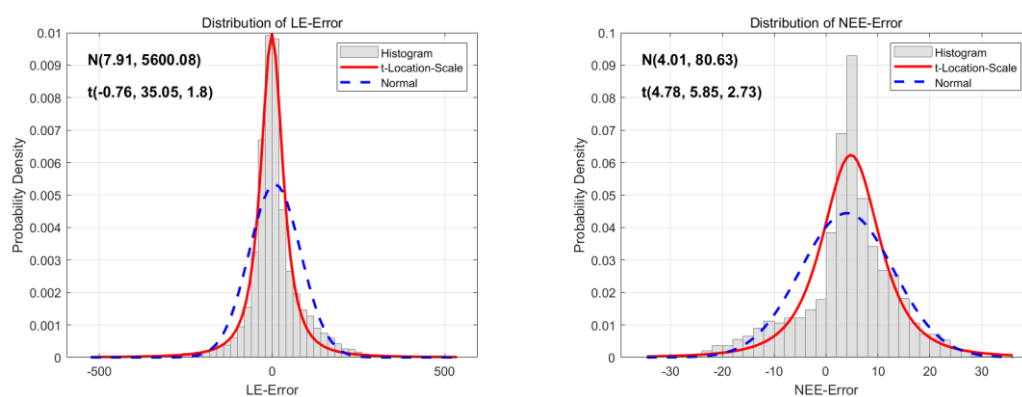


Figure. R1 Comparison of the fitting performance between the normal distribution and t-distribution for the residuals of the LE and NEE target variables.

6. – diagnostic metrics: also, could you please take the advantage of doing probabilistic

analysis to evaluate the posterior pdf using probabilitic metrics to look at reliability, precision and bias – ie not only deterministic (related) metrics, this only inspects one side of the history. Based on this, the advantages highlighted on the discussion section could be more defended.

**Response:**

Thank you for your valuable feedback.

We recognize that assessing the model's performance deterministically using the optimal solution from the posterior distribution (e.g., the posterior median) diverges from the main objective of uncertainty analysis in the Bayesian framework. To address this, the revised manuscript will include the full Bayesian predictive distribution based on the entire posterior distribution of the parameters, along with associated confidence intervals. Additionally, we will incorporate probabilistic metrics and uncertainty evaluation methods, such as those grounded in scoring rules, to provide a more comprehensive assessment of the model's fit. This approach will not only capture historical data but also fully utilize the advantages of Bayesian probabilistic analysis, thereby improving the model's interpretability and enhancing the reliability of the results.

7. – for me the text is a bit confusing when talking about multiple objectives, I guess most of the readers tend to think about multiple objective functions which is not the case bacause there is 1 likelihood – other thing is that there are 2 target variables.

**Response:**

Thank you for your feedback. The use of the term "multiple objectives" in our discussion may have led to the misunderstanding that we were referring to multiple objective functions. However, in this context, "multiple objectives" refers to the two target variables in the model (LE and NEE), not to multiple optimization functions. We employed a single likelihood function to fit these two target variables simultaneously.

In a deterministic parameter optimization framework, multiple objective functions are typically employed to find an optimal solution that minimizes the residuals of each target variable. However, in the Bayesian framework, we can represent the likelihood for multiple target variables as the product of their respective likelihood functions,

achieving joint optimization. This method allows us to account for uncertainty while simultaneously addressing multiple target variables, with each variable's contribution represented by its respective likelihood function. This approach effectively integrates information from different target variables, ensuring that the model captures not only historical data but also the inherent uncertainty in the parameters.

To prevent further confusion, we will clarify the distinction between "target variables" and "objective functions" in the revised manuscript, and provide a more detailed explanation of how a single (or joint) likelihood function is used within the Bayesian framework to model multiple target variables.

8. – maybe a naïve question, but do you need all the SA methods in the main manuscript?

**Response:**

Thanks for your question. In this study, we employed a two-stage approach that combines three qualitative sensitivity analysis methods (DT, MARS, and MOAT) with one quantitative method (Sobol') to identify the most influential parameters in the CoLM model for LE and NEE simulations. This approach was necessary to balance computational efficiency with robust parameter identification.

First, qualitative methods use relatively small sample sizes (hundreds to thousands) to provide an initial ranking of parameter influence by comparing model outputs across different input combinations. Given the complexity of the model, which involves nonlinearity and parameter interactions, relying on a single method risks overlooking key parameters. By combining multiple qualitative methods, we enhance the robustness of the initial parameter screening.

Second, while quantitative methods like Sobol' provide precise estimates of parameter contributions and interactions, they are computationally intensive, especially for high-dimensional parameter spaces. To mitigate this, we first applied qualitative methods to reduce the parameter set from 40 to the 10 most sensitive parameters (see Figure 1), thereby reducing the computational burden. In the second stage, Sobol' analysis was used to further evaluate these 10 parameters, identifying those that explained over 95% of the variance for further optimization (see Figure 2).

In summary, this two-stage process efficiently integrates qualitative and quantitative

methods, allowing for robust identification of key parameters while minimizing computational complexity.