General response: We appreciate reviewers' efforts for reviewing our revised manuscript. We revised the manuscript based on the following comments to further strengthen our manuscript. Please find our point-by-point responses in the following.

Reviewer #1
The authors have addressed most of my comments. I have some additional comments after reading the responses and revised manuscript.

-In the first-round review, a concern was about the better performance of the model with the default parameters than that with the optimal parameters. In the revised manuscript, the authors added an explanation in L 460-462 but dismissed quite quickly. Since the aim of model calibration is to obtain more reliable predictions (the focus of this work), the revision does not totally address the concern. The authors should elaborate the *potential deep causes* here.

Response: In this revision, we have now explicitly highlighted in the introduction section (Line 96) that the time scale for the calibration of ELM-simulated runoff is monthly. The optimal parameters thus obtained for calibration of monthly ELM-simulated runoff cannot guarantee an improved model performance at annual scale with respect to the default parameters. We also added explanations in the revised manuscript at Line 546 – Line 549.

-Please show the convergence curve of Gelman-Rubin R statistic in the manuscript.

Response: We added the convergence curve of Gelman-Rubin R statistic in the supplementary material Figure S6.

-The authors proposed to build a surrogate model for the RMSE metric to avoid constructing a surrogate for each grid. This is a good idea but a discussion on the potential limitation of this strategy would be helpful.

Response: We added two limitations in the revised manuscript. First, the surrogates of RMSE cannot be used to construct runoff uncertainty, therefore, ELM simulations are still needed. Another limitation is using RMSE at monthly scale as objective cannot guarantee the performance at annual scale. Please find details in Line 542 to Line 549.

Reviewer #2
I appreciate the authors taking the effort to address my comments, particularly related to adding a new period of simulation to test the calibrated model. While the revision has improved the manuscript, it does not address some of my comments, re-iterated below. It seems that some necessary (and relatively minor) changes are needed before the manuscript can be accepted.

- The novelty is still unclear to me. The authors clarified in the revision that "The selection of RMSE as QoI in constructing surrogate models is a novelty of this work, which can significantly reduce the computational burden of surrogates' construction and parameter inference." (line 213-214 in the manuscript with tracked changes). As pointed out in the previous review comments, this (using RMSE as QoI) is not new (e.g. Wang et al. (2014); Razavi et al. (2012) and references therein). Developing a surrogate model for a performance measure (RMSE in this case) and then optimizing it is actually the focus of Razavi et al. (2012) paper cited in the manuscript, and these methods have had a good number of hydrologic applications in the last decades. Therefore, I would suggest the authors highlighting their contribution in the analysis results such as runoff patterns before/after calibration, rather than the surrogate modeling method itself.

Response: We reworded the sentences that highlight the selection of RMSE as QoI is the novelty in the revised manuscript as listed in the following.

In abstract, we modified "The main methodological advance is this work is the selection of error metric between the ELM simulations and the benchmark data is selected to construct the surrogates, which facilitates efficient calibration and avoids the more conventional, but challenging, construction of high-dimensional surrogates for the ELM simulated runoff." to "Error metric between the ELM simulations and the benchmark data is selected to construct the surrogates, which facilitates efficient calibration and avoids the more conventional, but challenging, construction of high-dimensional surrogates for the ELM simulated runoff."

In the method section, we modified "The selection of RMSE as QoI in constructing surrogate models is a novelty of this work, which significantly reduce the computational burden of surrogates' construction and parameter inference." to "The selection of RMSE as QoI in constructing surrogate models significantly reduce the computational burden of surrogates' construction and parameter inference."

- Thank you for adding information regarding the convergence of the MCMC chains. Please further clarify at what stage the reported Gelman-Rubin R statistics were evaluated, i.e., were they calculated at the 1,000th iteration (the end of burn-in), or the 10,000th sample. Only samples after convergence can be retained to form the posterior distribution, so the R statistics at the 1,000th iteration is needed here.

Response: The Gelman-Rubin R statistics were estimated with the samples after the burn-in period. We have now clarified it at Line 371.

- Likelihood function Eq. (21) - I don't follow the added text "where sigma is estimated as the standard deviation of RMSEs between simulated runoff and GRUN from all the training simulations because the objective is to minimize RMSE". If I understand correctly, sigma should refer to the standard deviation of the error (difference between GRUN runoff and simulated runoff), different from the text. For example, RMSE is always positive, but error can have both signs.

Response: Yes, the reviewer is right that σ refers to the standard deviation of the error. In Equation (18), the error represents the difference between GRUN runoff and simulated runoff when runoff is the QoI. However, since we used the RMSE as QoI for constructing surrogate, Eq (18) is transformed to Eq (21) when evaluating likelihood in the parameter inference process:

$$\log L(\boldsymbol{y}|\boldsymbol{X}) = -\frac{N \cdot (0 - RMSE^{PC})^2}{2\sigma^2} - \frac{N}{2}\log(2\pi\sigma^2)$$

Here, we assume σ refer to the standard deviation of difference between RMSE and 0 because RMSE is our QoI and 0 represents RMSE of observation. In other words, the objective is to find the parameter with RMSE as close to 0 as possible. So, we estimated σ as the standard deviation of RMSE between simulated runoff and GRUN from all the training simulations. In this revision, we modify Eq (21) to show RMSE = 0 is the objective, and clarify σ has a different meaning than previous equation (Line 223 – Line 225).

- Following up on the other reviewer's comment - since the optimal parameters fit the monthly runoff better than default but not the annual, I suggest replotting Fig. 11 to show monthly runoff and discussing the calibration gains on capturing the monthly runoff patterns.

Response: Apart from calibrating the ELM-simulated runoff at monthly scale, the study also aims to quantify the parametric uncertainty of the simulated runoff and understand how the parametric uncertainty impacts annual runoff trend. Therefore, in Sec 5.7, we present analysis of the annual runoff to illustrate the parametric uncertainty and perform trend analysis. The Figure 11 shows the parametric uncertainty of the annual runoff at basin scale is significantly reduced after parameter inference. Additionally, plotting Figure 11 at monthly scale will make the figure very busy, thus affects the readability.