We would like to thank the reviewer for their comments. We included the original comment in black font and **our response in bold violet font**.

Any planned changes or additions to the text are in violet font with boxes around them.

Response RC1 – Zeli Tan

Feldbauer et al. present a timely and very interesting study related to the uncertainty of physical lake models. Given the critical role of the studied models in climate impact analysis, such as ISIMIP, this research provides important implications for a better assessment of global change in lakes and reservoirs. The manuscript is with high quality of presentation and rigorous scientific inquiry. Only some clarifications and extended discussions are needed before it can be accepted for publication.

We thank the reviewer for the positive words and we reply to the specific comments further down.

One limitation of the current approach, which simultaneously assesses three types of uncertainties in lake models (i.e., input uncertainty, parameter uncertainty, and structural uncertainty), should be further discussed. Due to the interactions among these uncertainties, it is challenging for this multi-objective approach to fully resolve specific uncertainties. As admitted in the manuscript, the uncertainty of the three input-related scaling factors may hide the uncertainty of model-specific parameters and process descriptions. Consequently, the method obscures the investigation of an optimal model for specific lake types and an optimal algorithm for specific lake processes. Notably, for global lake modeling, we hope that as climate data become more accurate with time, the uncertainty of global lake simulations will be reduced. We also hope that lake models can achieve consistent global simulations without lake-specific calibrations. Despite the great values of the current paper, it falls short in addressing these issues. Conversely, I would encourage the authors to conduct a future study which uses observed atmospheric forcings to exclude the uncertainty in input data and ensure the good model performance achieved for right reasons. There are some existing studies following the recommended approach, such as Guseva et al. (2020) and Guo et al. (2021). But their values are limited due to either focusing on only one lake or testing only one lake model.

Guo, M., Zhuang, Q., Yao, H., Golub, M., Leung, L. R., & Tan, Z. (2021). Intercomparison of thermal regime algorithms in 1-D lake models. Water Resources Research, 57, e2020WR028776. https://doi. org/10.1029/2020WR028776

Guseva, S., Bleninger, T., Jöhnk, K., Polli, B. A., Tan, Z., Thiery, W., ... & Stepanenko, V. (2020). Multimodel simulation of vertical gas transfer in a temperate lake. Hydrology and Earth System Sciences, 24(2), 697-715.

We agree with the overall statements that the reviewer makes, and that looking at multiple aspects of uncertainty makes it harder to focus on each individual contribution. Our reason for including input scaling factors was the considerable uncertainty that is present in these variables (i.e., meteorology and water

transparency) for global simulations. The reviewer suggests an alternative approach for a future study. We wholeheartedly support such an idea, but this is currently infeasible with the ISIMIP data. As local meteorological forcing was not supplied and is likely not available for all sites, the bias-adjusted reanalysis data used here was the best available option for modeling. The proposed approach by the reviewer would therefore require a considerable data collection effort. Though we agree with the reviewer that this could be worth it, as observed meteorological data strongly reduces input uncertainty. We would raise the point, however, that even locally-observed meteorological forcing does not fully exclude uncertainty about the input data, as 1D models integrate signals from the entire lake and even meteorological observations in the center of the lake may not be representative of what the whole lake experiences. This is especially true for wind speed, which can have large temporal and spatial variations.

In L. 356-357, we already outlined the added benefit of using local observations. We want to modify this statement to incorporate some of the reviewer's suggestions:

In a setting with locally observed meteorological forcing data, the model-specific parameters might become more influential, if meteorological forcing variables can be more constrained. Previous studies used this approach in one or a few lakes (e.g. Guseva et al. (2020); Guo et al. (2021)), but it would be beneficial to compile such data for a larger number of lakes, similar to the present study. Reducing the strong influence of meteorological scaling factors could facilitate identification of optimal models for different clusters. If observations are not available, improvements in downscaling methods from global products to weather conditions at the lake surface might also partially achieve this.

We will add a section to the discussion highlighting the reviewer's main remark:

The overall uncertainty of mechanistic simulations is usually related to uncertainty in the initial conditions, uncertainty in the driving data (both forcing data such as meteorology and data used for calibration such as water temperature), uncertainty in the model parameter values, and structural uncertainty in the process description also called epistemic uncertainty (Thomas et al. 2020, Scavia et al. 2021, Dietze 2017). In this study we wanted to explore the relationships between lake model performance, parametrization, and lake characteristics, so we are mainly concerned with the uncertainties related to parameter values and model structure. The uncertainty in the meteorological forcing is thereby partly acknowledged by the inclusion of the scaling factors. Because the scaling factors proved to be amongst the most sensitive parameters, they could potentially prevent identification of an optimal model or patterns relating the parametrization of the models to the lake characteristics, if such an optimal fit exists. A way forward could be to reduce the uncertainty in the meteorological forcing data, and hence hopefully the sensitivity of the scaling factors, by using local meteorological observations instead of reanalysis data.

Literature:

Guseva, S., Bleninger, T., Jöhnk, K., Polli, B. A., Tan, Z., Thiery, W., Zhuang, Q., Rusak, J. A., Yao, H., Lorke, A., & Stepanenko, V. (2020). Multimodel simulation of vertical gas transfer in a temperate lake. Hydrology and Earth System Sciences, 24(2), 697–715. https://doi.org/10.5194/hess-24-697-2020

Scavia, D., Wang, Y.-C., Obenour, D. R., Apostel, A., Basile, S. J., Kalcic, M. M., Kirchhoff, C. J., Miralha, L., Muenich, R. L., & Steiner, A. L. (2021). Quantifying uncertainty cascading from climate, watershed, and lake models in harmful algal bloom predictions. Science of The Total Environment, 759. https://doi.org/10.1016/j.scitotenv.2020.143487

Thomas, R. Q., Figueiredo, R. J., Daneshmand, V., Bookout, B. J., Puckett, L. K., & Carey, C. C. (2020). A Near-Term Iterative Forecasting System Successfully Predicts Reservoir Hydrodynamics and Partitions Uncertainty in Real Time. Water Resources Research, 56(11). https://doi.org/10.1029/2019WR026138

I suggest the authors to avoid the use of "hydrodynamic lake models" to describe the studied models. To me, hydrodynamic models refer to numerical models that solve the transport of both mass and momentum. The authors can use "physical lake models", "thermodynamic lake models", or just "lake models".

It is up for debate whether the processes described by these one-dimensional models are part of “hydrodynamics” or not; for GLM, GOTM, and Simstrat, we would argue this is the case, but we acknowledge that the FLake model simplifies many hydrodynamic processes. To avoid confusion about this terminology, we will refer to all models as “(process-based) lake temperature models”.

We will change the usage of the term in the manuscript accordingly throughout the manuscript.

In the methodology, one area that needs clarification is what procedure the authors have adopted to ensure appropriate initial conditions for simulations. It can be particularly important for the modeling of deep lakes.

Yes, we had previously not provided this information. To give more information, we will append Section 2.1 with the following paragraph:

Initial conditions were estimated from observed water temperatures. Therefore, all available data in a period of days (depending on data availability) before and after the start date of the simulation were taken and averaged to set the initial temperature profile. All simulations used a spin-up period of 1 year.

The scripts to set up the calibration are linked to in the Code and data availability section of the manuscript (https://zenodo.org/doi/10.5281/zenodo.13165427).

Minor comments:

L55: also Zhuang et al. (2023). Zhuang, Q., Guo, M., Melack, J. M., Lan, X., Tan, Z., Oh, Y., & Leung, L. R. (2023). Current and future global lake methane emissions: A process-based modeling analysis. Journal of Geophysical Research: Biogeosciences, 128, e2022JG007137. https://doi. org/10.1029/2022JG007137

We thank the reviewer for this relevant reference and we will add it to the examples.

L181: gets absorbed

We will revise this.

L202: How does the metric δ differ from the Sobol's total-order index?

Delta moment-independent measure should not be classified as a total-order index. We plan to correct this throughout the manuscript. Delta (δ) represents the importance of the entire distribution of the specific model parameter with respect to the entire distribution of simulated water temperatures (Plischke et al., 2013; Borgonovo, 2007; both cited in the main text). In contrast, Sobol' S1 estimates a parameter's influence on the variance of the simulated water temperatures. In this analysis, delta measures provide a second estimate to complement the variance-based Sobol first-order index to strengthen the analysis. As this study is interested in identifying the most important parameters (i.e. factor prioritization setting), we follow the recommendations of Borgonovo et al. (2017; cited in main text) and use both variance-based and moment-independent measures. We will change the description of both measures (from line 202):

> The analysis calculates two sensitivity measures, the moment-independent δ and variance-based Sobol S1. The delta moment-independent measure δ considers the influence of the entire distribution of a model parameter with respect to the entire distribution of simulated model output, whereas the variance-based first-order sobol index S1 calculates a parameter influence on the variance of the simulated model output (Plischke et al., 2013; Borgonovo, 2007). As this study was interested in identifying the most important parameters (i.e. factor prioritization setting), we followed the recommendations of Borgonovo et al. (2017) and used both variance-based and moment-independent measures to increase the robustness when estimating the inference of which parameters are most important when simulating water temperatures.

And from line 251:

> From the calibration runs using the latin hypercube approach, we calculated moment-independent measure δ and variance-based first-order measure S1 for each combination of models, performance metrics, and lakes (Figure 5).

To take into account parameter interactions, we also calculate S_interactions and discuss its importance for simulating water temperatures.

L212: What is "SA"?

We meant "sensitivity analysis", but this was an oversight and we will now write "sensitivity metrics" instead.

L226-227: I suggest moving Figure 3 upward to Section 2.2

We will follow this suggestion, as this would give readers an overview of the lakes we simulated early on in the text.

L232: Figure S5 is introduced prior to that of Figure S4.

We thank the reviewer for noticing this and we will switch the order of the two figures.

Figure 5: It is surprising to see that S1 is larger than δ in many cases. To my experience, the first-order sensitivity should be smaller than the total-order sensitivity.

We appreciate the reviewer's keen eye and the opportunity to strengthen our description of the sensitivity analysis metrics used. Please see our comments and revisions above (to the comment on delta and Sobol measures). In addition, it is not uncommon to see Sobol S1 values larger than delta values in the sensitivity analysis literature.

L306: remove "a"

We will revise this.

L407: potentially

We will revise this.