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 RC2 – Fabian Bärenbold

General comments

One-dimensional physical lake models are a widely used tool in simulations of lakes and reservoirs for diverse goals like now-casting, forecasting, or to compute mixing for use biogeochemistry. Although many different models exist and several of them are very widely used, to my knowledge no consistent comparison between them exists until now on a wide range of lakes. In general, research about the link between lake types, model parameters and uncertainty is not widespread and I think this manuscript is a good contribution to fill this gap. The manuscript is well organized and written with only few spelling mistakes. However, I think that there is a problem with one of the calibrated parameters of GOTM and not enough details on the observational data.

We thank the reviewer for the positive feedback and helpful comments. We address the concerns below.

Specific comments:

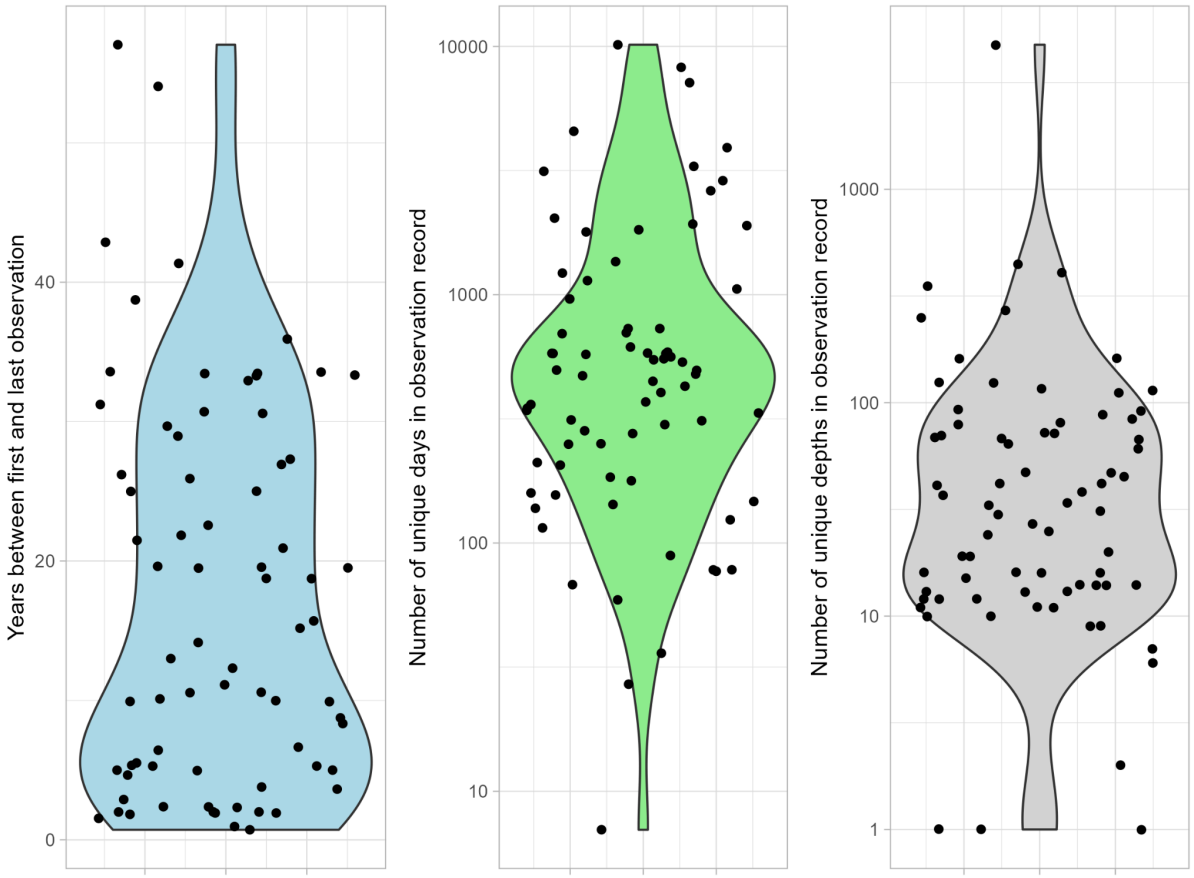
The observational data used to calibrate the 73 lake models is of great importance, but very few details are given in this publication. I would find it useful to a small paragraph about time and depth resolution of the observational data and whether any weighting was done to compute performance metrics (lines 196 – 198). Is there a minimum requirement of observations to be included to the 73 lakes? In addition, I would welcome a Figure giving some information about depth and time resolution of measurements in the supporting information.

We acknowledge that we provided too little information about the observational data. We will modify lines 64-65 to:

The resolution (vertical and temporal) of the observed data and the detail of the hypsograph varied per lake. For all but two lakes, data covered a period of at least 1 year, and for 75% of the cases it covered at least 5 years. Profiles (three unique depths or more) were provided for all but four lakes, and all lakes had more than 100 unique observations (Figure <X> in the Supplementary material). A link to the observed data and hypsographs is provided in the Code and data availability statement.

The available data varied quite a lot, so there would often be one or two exceptions to general statements about the data. We will provide a link to the observed data, as made available to ISIMIP modelers, in the Code and Data availability statement: https://github.com/icra/ISIMIP_Local_Lakes/tree/main/LocalLakes.

We agree with your suggestion and we will add a figure to the Supplement:



The very high number of unique depths in observation records are caused by the fact that some of the observational data were from CTD profiles, which we aggregated to a vertical resolution of 0.1 m before plotting. This information will be included in the figure text.

Weighting was not performed. We will modify L. 196-198 to state that the metrics were calculated on all water temperature observations:

> We sampled and ran the four models for 2000 parameter sets, and for each of the parameter sets we calculated four performance metrics over all water temperature observations: Root mean squared error (RMSE), Nash-Sutcliffe model efficiency (NSE), Pearson correlation coefficient (R), and mean error (bias)

Is there a reason the GSWP3-W5E5 reanalysis dataset was chosen? Could you explain this in 1 – 2 sentences? Also a bit related, is there any chance of going to hourly instead of daily resolution? As far as I understand from the discussion this might solve some of the problems (wind speed effect on mixing is cubic). If yes, it would be interesting to mention this in the discussion/conclusion.

The GSWP3-W5E5 dataset was chosen because one of the aims of the calibration was to run the ISIMIP3 local lakes climate simulations, and W5E5 was used for the bias-correction in ISIMIP3. These climate simulations are not part of this paper, but can be found on the ISIMIP data portal (linked to in the Code and data availability statement). We now briefly mention the aim of the calibration in the text, but we reject

adding a long explanation in order to not confuse the reader by referring to simulations that are not part of the paper.

After L. 45, we plan to add:

> These calibrations were in preparation for ISIMIP climate impact simulations for the local lakes sector (see Code and data availability).

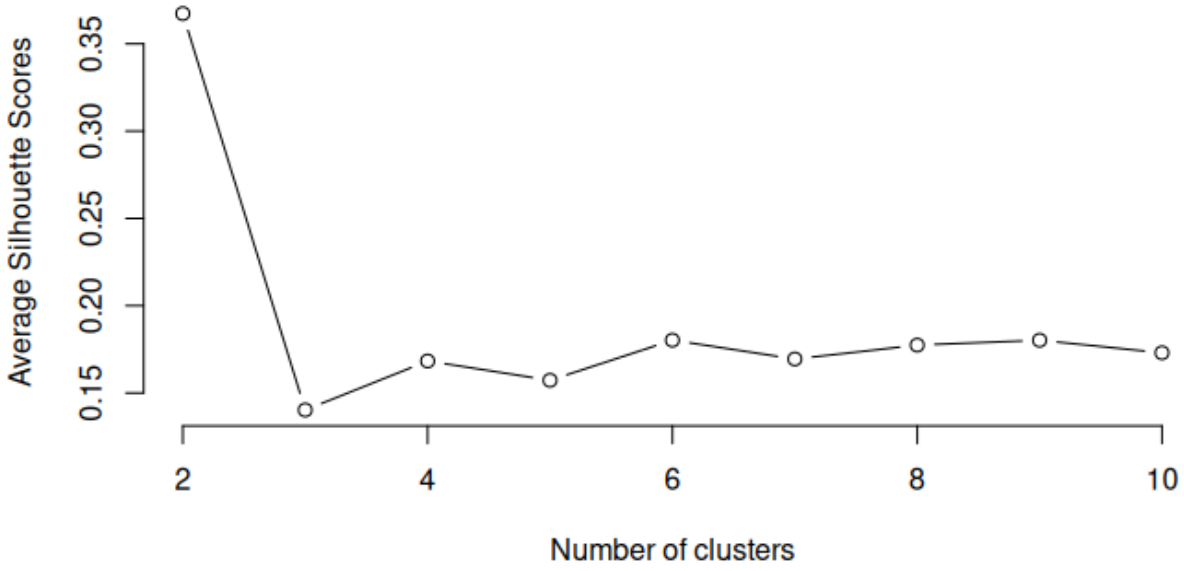
The effects of going from daily to hourly resolution are interesting. To our knowledge, there are no hourly meteorological forcing data available in ISIMIP3 at the time of writing. However, temporal downscaling could be considered to take into account the importance of diel variations, as the reviewer is suggesting. This was done in a previous study by Ayala et al. (2020, cited in main text) using local observations and artificial neural networks. Their results suggested indeed an improvement when using hourly data compared to daily, although improvements when using synthetically-generated hourly data were minor. Relating to the reviewer's comment about wind speed: their calibrated wind scaling factor was similar (1.5) when using daily or synthetic hourly, while it was lower when using observed hourly forcing (1.4). We expect that when focusing on general physical trends and long-term changes in water temperature, daily forcing is sufficient. However, when simulating short-term events or extending to biogeochemical simulations, sub-hourly forcing will become more valuable or even necessary.

To mention the effect of a higher temporal resolution in the text, we plan to append the text after L. 357:

> Similarly, use of hourly meteorological forcing could result in more realistic patterns in wind-driven or convective mixing (Ayala et al., 2020).

I would find it interesting to have the silhouette plot of the cluster analysis in the supporting information.

We plan to append the silhouette plot to the supporting information. As stated in the manuscript we manually chose to use 5 lake clusters even though the silhouette plot suggested the optimum number was 2, because we are convinced that it is more informative. To arrive at the 5 clusters we tried out different numbers of clusters (4, 5, 6) and decided that 5 was the most informative (i.e., containing most scientifically interesting limnological characteristics for discussion) while not creating too many clusters with very few members.



The authors discuss the fact that k_min of GOTM seems to be the most sensitive of the lake-specific calibration parameters. The range of k_min used in this study is $1.5e^{-7}$ to $1e^{-5}$, which is rather high values compared to a default value of $1e^{-8}$ in the GOTM manual. In addition, typical values of ~$1e^{-6}$ are reported for TKE in the hypolimnion of lakes (e.g. Wüest and Lorke, 2003). I see a major problem if k_min is set too high: it could, together with high values for the scaling of shortwave radiation, offset a low value for the scaling of wind speed. There is evidence of this in Figure 7, where the calibrated parameters for GOTM are consistently low (wind speed) or high (shortwave) compared to the other models. The calibrated value of k_min always seems to be well above $1e^{-6}$ (Figure S8), so on the order of typically observed values for TKE (Wüest and Lorke, 2003). The authors discuss the potential interactions between k_min and wind scaling on one hand (lines 341 – 345), and shortwave scaling and wind scaling on the other hand (lines 365 – 369) but not the potential interaction between all 3. In regard of this, I would like to ask the authors to motivate their choice of the range of k_min and to check its influence on the wind and shortwave scaling parameters. I see two ways to do this:

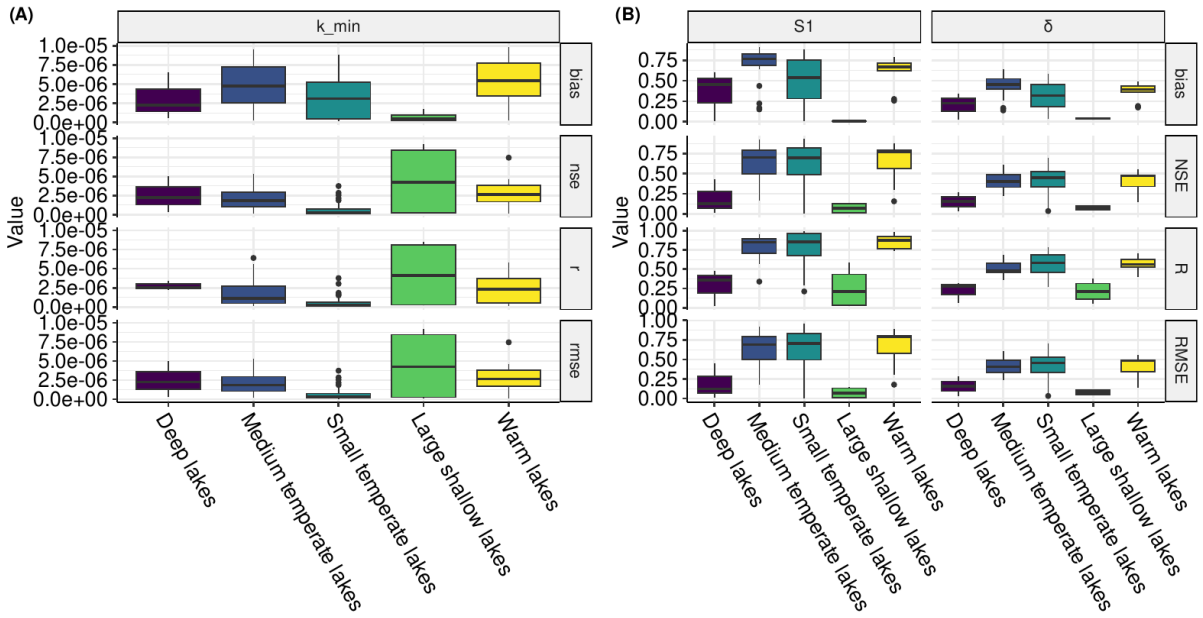
- Compute second/third order Sobol indices for the concerned parameters
- Redo some of the simulations with k_min = 1e-8

I understand that the interaction calculation shows that this suggested parameter interaction is not driving variance but the correlation could still be strong among these 3 parameters. There is also no obvious reason why wind and shortwave scaling should be so different for similar models like Simstrat and GOTM. I could be wrong but to me it seems like k_min in GOTM is playing the role of the seiche in Simstrat.

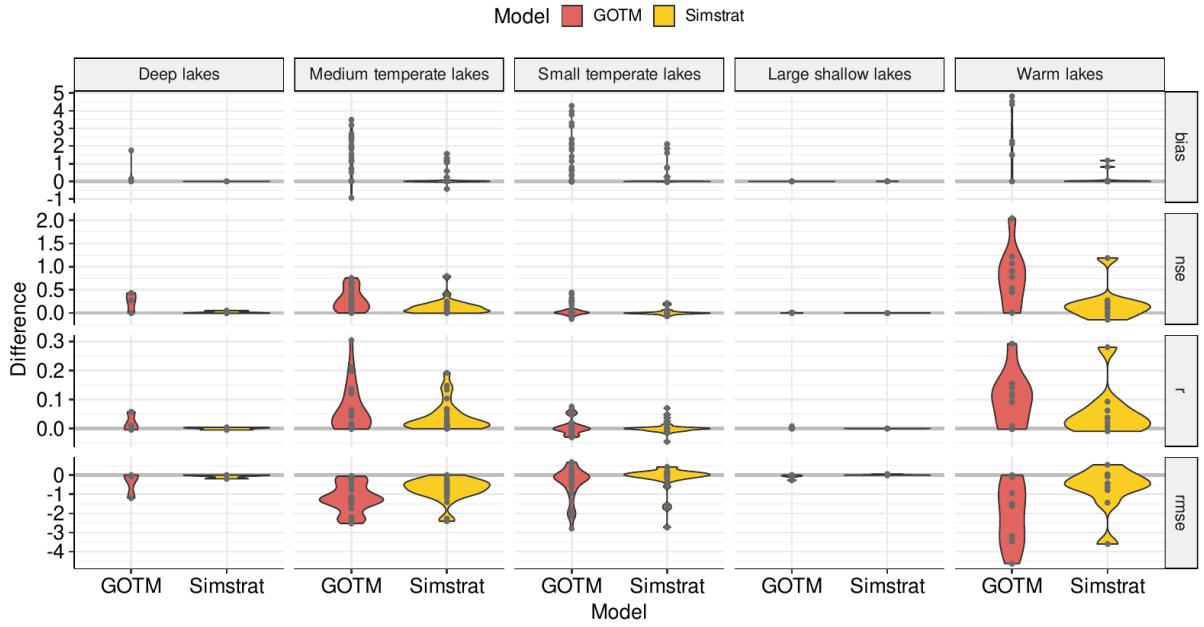
We chose the range for the calibration of k_min based on past experience and discussion with colleagues experienced in GOTM (e.g. Andersen et al. (2020) & Ayala et al. (2020); both cited in the main text). We believe that the default value for k_min from the GOTM manual is for the ocean as we are not aware of a reference manual for

the lake branch of GOTM. As mentioned in the reviewers comment, typical values for TKE in lakes are larger. We got feedback from the GOTM developers that: "k_min is 'un-resolved TKE generation' and should in principle cover seiching. You write that k_min value varies a lot between lakes, which could be a result of the importance of seiching for a given lake." (Personal communication with Karsten Bolding, 2024).

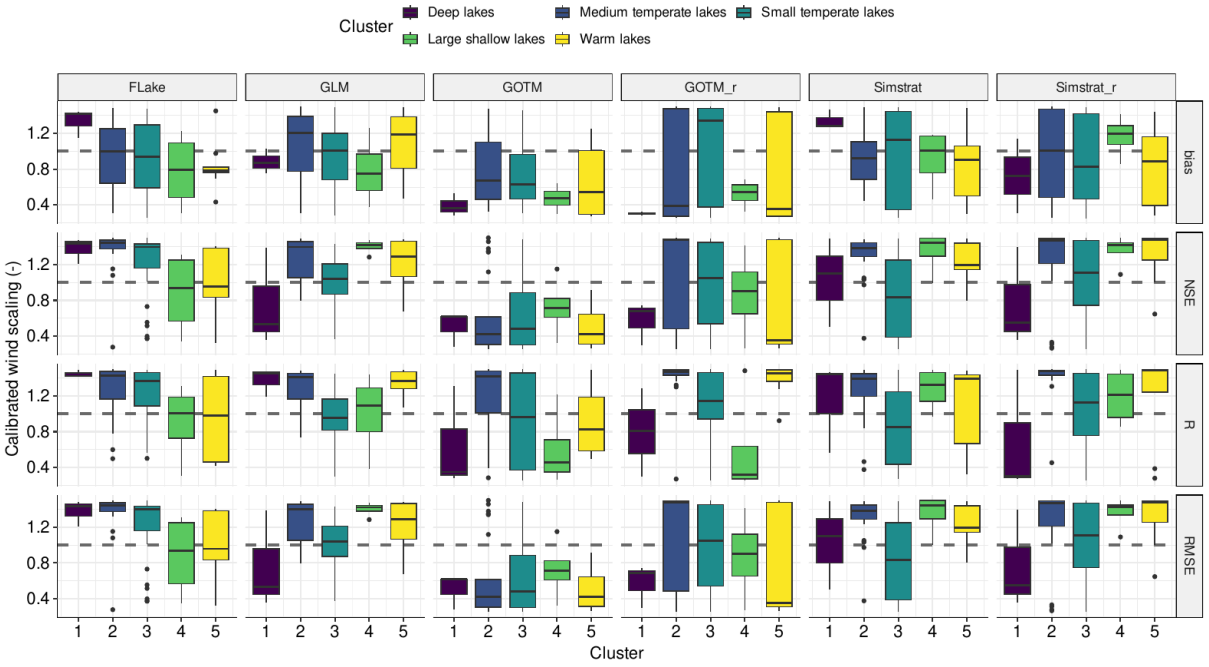
We further investigated the calibration results and found that the k_min values for the best performing parameter set are larger for deep, medium temperate, and warm lakes compared to small temperate lakes (see attached figure below, subpanel A). The exceptions are the large shallow lakes that cover a wide range of k_min values. This could be explained by looking at the sensitivity of k_min for the different clusters, whereas for the large shallow lakes k_min has very low sensitivity values (Figure below, subpanel B).



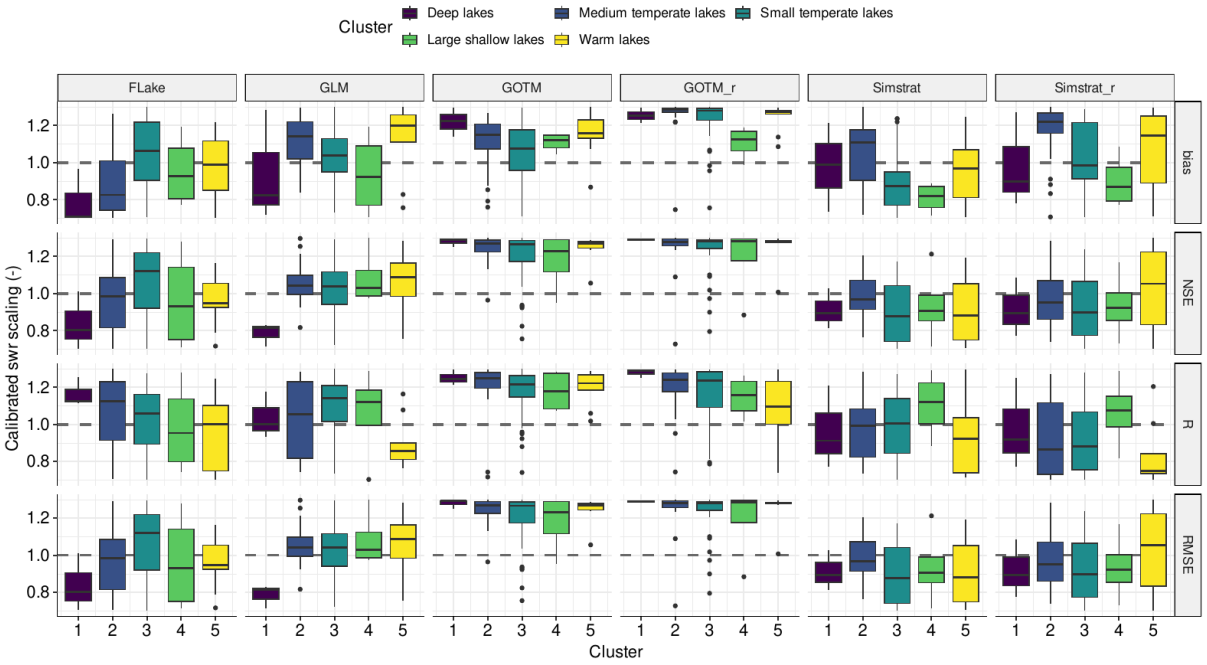
We do agree with your comment regarding potential interactions or correlations between swr scaling, wind speed scaling, and k_min. So we followed your suggestion and ran an alternative round of calibration for GOTM with k_min fixed at 1e-8. In addition, we ran an additional round of calibration for Simstrat with a_seiche set to 0, which disables the seiching module for Simstrat. This was suggested by a colleague to test if the improved performance of Simstrat was caused by the inclusion of seiching as we speculate in the discussion and allows us to better compare the impact of reducing k_min (which as mentioned above should also cover seiching). Under these additional calibration rounds (GOTM with k_min = 1e-8 and Simstrat with a_seiche = 0) both of the models showed worse performance in most of the lakes, especially for the medium temperate and warm lakes. For large shallow lakes, very small differences were seen and for some of the small temperate lakes we even saw better performance for both Simstrat and GOTM (See attached figure below; the difference in the plot is calculated as old calibration - new calibration).



However, as the reviewer suspected, with k_min kept at 1e-8 the calibrated wind speed scaling from GOTM (GOTM_r in the attached figure below) increased and became more similar to the other models.



For the calibrated swr scaling we do not see a reduction in the values for the best performing parameter set for the new round of calibration (GOTM_r in the attached figure below). The calibrated values even slightly increase when keeping k_min at a constant value of 1e-8.



Reducing k_min reduced the overall model performance of GOTM for the lakes where deep mixing (incl. internal oscillations) is of importance (as seen by the similar reduction in model performance of Simstrat when a_seiche = 0). In GOTM, increasing wind speed scaling factor can (to some degree) compensate for this, but it is not able to perform nearly as well as with larger k_min values. As GOTM cannot reach similar performance by increasing wind speed scaling and reducing swr scaling, we suggest that there is potentially no strong interaction or correlation between the three parameters (as previously seen by the sensitivity and interaction measure). Only for some of the small temperate lakes, a lower k_min value (and a_seiche = 0 for Simstrat) is resulting in better model performance.

We will add a paragraph reflecting this discussion in the supplementary information. Also, we plan to update all scripts and add the new calibration data to our github and the Zenodo repository

In L324 we plan to add

> We reinforced this hypothesis by performing additional simulations with a_seiche set to 0, which lead to poorer model performance of Simstrat (see supplementary material for details).

In L340 we plan to add

> for all lake clusters besides the large shallow lakes (Figure <Y> in the Supplementary material).

In L386 we plan to add

(more details on this can be found in the supplementary material).

The authors seem to imply that larger lakes should generally have a larger value for wind scaling and vice versa for small lakes (lines 349 – 351). However, if true this effect should be visible in the calibrated wind scaling parameters, no?

The mentioned implication was our initial hypothesis, but our study did not find a relation between best parameter values for wind speed scaling factor and lake size to confirm this. A possible reason for this is the resolution of the gridded data. In order to clarify this, we plan to add a sentence:

We could not highlight any relations between best parameter values for wind speed scaling factors and lake size, which could imply the gridded weather data mask any effects of lake size.

Technical comments:

Title: maybe “hydrodynamic” or “physical” lake models instead of just lake models

We agree this title was not restrictive enough. We will modify the title (also in connection to RC1’s comment) to:

Learning from a large-scale calibration effort of multiple lake temperature models

Line 6: The following sentence is a bit too unspecific to me: “The models performance and parameter sensitivity showed a relation to the lake characteristics and model structure.”

We agree and will modify the sentence accordingly:

Parameter values, model performance, and parameter sensitivity differed between lake models and between clusters that were defined based on lake characteristics.

Line 39 - 40: Maybe mention that although important for shallow lakes, the biogeochemical components are not discussed in this manuscript.

We agree that it was unclear if we were talking about our study or Andersen et al. (2021). We will modify L. 37-40 as follows:

Andersen et al. (2021) performed an extensive, global sensitivity analysis on the 1D coupled physical-biogeochemical model GOTM-WET in three Danish lakes and found that parameter sensitivity may be strongly linked to lake morphology, including a potential

feedback of biogeochemical components on temperature (such as light absorption by organic matter) in shallow lakes.

Line 64: Make clear that Table S1 of the current manuscript is meant and not Table S1 of Golub et al. (2022)

We will change this to make it more clear to the reader by citing the Golub et al. paper in the previous line.

Line 82: delete “and” before “the 1D turbulence-based models GOTM”

We will revise this.

Line 99: “comma” after “Equation 1”

We will revise this.

Equation 3: Shouldn’t it be h < z < D_lake?

Yes, thank you for noticing this. We will revise it.

Equation 5: Please check whether term 1 is really negative.

Yes, the first term on the right-hand side needs to be negative to account for the vertical gradient of short–wave radiation to act as a heat source for the water column with the underlying assumption that the depth dimension is positive (reference at the surface).

Line 220: This sentence is not clear to me. Did the authors intend to say “between best and worst performing model”?

Yes that is what we wanted to say. This will be revised.

Line 263: “were” instead of “was”

We will correct this.

Line 269: “Lake clusters” instead of “lake cluster”

We will correct this.

Line 357: “better” instead of “more”

We will correct this.