Response to anonymous reviewer #2.

The authors are grateful to the reviewer for his attentive reading of the paper and for pointing out some statements in the paper that seem inconsistent. Addressing these issues allowed us to improve the presentation of the results. The specific comments are considered below.

General comments

Reviewer: *"Pag. 4003;L3-5: How were the initial conditions prescribed for the simulations/models with active bottom sediments ?"*

Authors: There was no convention on initializing soil temperature profile. We inserted a clarification in the "Experimental setup" section.

R: *"Pag. 4007:L3-6: The results in Fig. 3 do not seem to support this statement pointing to the cold bias due to neglecting bottom sediments by SimStrat and LAKEoneD. CLM4 and MINLAKE96 also have a cold bias in the second period and they represents bottom sediments."*

A: We attribute the cold bias in SimStrat and LAKEoneD during autumn to the lack of heat flux from bottom sediments because LAKE, which belongs to the same category of models (k-epsilon turbulent closure models), explicitly accounts for sediments and does not demonstrate a cold bias. Note also that during summer, when the heat flux from the sediments does not significantly affect the surface temperature (see Section 5.1.2), the biases computed for the k-epsilon models are very similar. Cold biases of CLM4 and MINLAKE96 may be caused by a number of factors, e.g. soil temperature initialization, soil properties and soil depth among others. We do not address the possible reasons as that would shift the presentation from the focus of the paper.

R: *"Pag. 4009:L10-19: These results seem to contradict the last sentence of the abstract: "while the effect of heat flux to bottom sediments can become significant for bottom temperatures, it has no important influence on the surface temperature". However, in this section it is estimated that "neglecting sediments may account for at least 50% of the seasonal surface temperature difference error in k-e models". Please clarify these statements."*

A: We modify the sentence in the Abstract as follows: "While the effect of heat flux to bottom sediments can become significant for bottom temperatures, it has no important influence on the surface temperatures in summer."

R: *"Pag: 4010:L26:26: "These results also suggest that the "optimal depth" delivering the most realistic surface temperature is model-dependent": This sentence seems to contradict the interpretation of table 4 (pag 4010:L23:34 "The data of Table*

4 do not allow a conclusion…"). Looking at whole period in Table 4, the lower DM (absolute) is for the simulations with the local depth in all models. However, the differences between the different simulations are very small (and most likely statistically insignificant). Therefore, these results would suggest that for these types of shallow lakes there is a reduced sensitivity of lake depth and an "optimal depth" derived from a single model could be applied to other models."

A: We agree with the reviewer that the interpretation of the results of this section was not precise enough. Thus we embedded the text at the end of Section 5.1.3 "These results also suggest that the ``optimal depth" delivering the most realistic surface temperature is model-dependent. However, due to the low sensitivity of surface temperature to variations in lake depth around the mean depth (where the optimal depth is likely to fall in all models), an optimal depth delivered be one lake model will not cause significant errors in surface temperature when used in other models. Therefore, the global lake depth datasets derived by minimizing the surface temperature error of a particular model (e.g. Balsamo et al., 2010) can be used in NWP and climate models utilizing other lake parameterizations (at least for relatively shallow lakes)."

R: *Pag. 4011:L26:28: "During autumn FLake and Hostetler, along with other models, reproduced well the almost homogeneous thermal structure (weak stratification) developed due to convection (not shown)." : These results are presented in table 5, where Flake and Hostetler models have DM in the second period (for the reference run) of 1.36 and 1.56, respectively. These systematic errors are quite large when compared with the mean observations of 0.26. On the other hand, LAKE shows reduced DM. These results suggest that FLake and Hostetler models still have a strong stratification in the second period.*

A: We corrected the sentence as follows: "During October and beginning of November, FLake and Hostetler models, along with other models, were able to reproduce the almost homogeneous thermal structure (weak stratification) developed due to convection (Fig.~4)." Indeed, large DMs of 0-1 m temperature difference in autumn in these models are caused by overestimated stratification in August and September (Fig.4)

R: *Different models used in different experiments: Along section 5 the results of different experiments do not include all the models (e.g. zero heat flux :MINLAKE96 is missing (table 3); different lake depths: only 3 models (table 4)). This makes the discussion sometime confusing. Not all the models performed all the sensitivity simulations? If this was the case, I understand the doing those extra simulations can be time consuming, but it would make the discussion of the results more clear. I leave this suggestion to the author's consideration.*

A: Due to organizational and technical issues, all the models could not be run through all experiment configurations. However, we still believe that results from experiments

with a reduced number of “participants” clearly identified some effects of model performance that can be anticipated in other models.

R: *Abstract: The abstract mentions the turbulent fluxes error between 17-28 Wm2, but is should also include that the residuals of the heat balance, based on observations, fall in the same range.*

A: We modified the Abstract as follows: “Total heat turbulent fluxes, computed by the surface flux schemes of the compared lake models, deviated on average from those measured by eddy covariance by 17—28 W/m^2. There is a number of possible reasons for these deviations, but the conclusion is drawn that underestimation of fluxes is very likely to be due to eddy covariance technique.”

Minor changes

We adopted all recommendations of the reviewer, excluding the one:

R: *Pag.4005: The use of the term “DM”, for the differences of modelled and observed means is not very common. I would suggest using systematic bias, or just bias.*

A: Although the term “bias” is widely used to indicate the difference of means of two time series we avoid using it because it does not clearly correspond to strict definitions applied in statistics. For instance, in statistics bias may stand for the difference between an estimator's expectations and the true value of the parameter being estimated. However, in our case measured values are not always may be considered as “true values”, as it is apparently so for eddy covariance measurements.