Response to anonymous reviewer #1.

The authors are grateful to the reviewer for providing helpful remarks, which allowed us to improve the presentation of the results, and for bringing to our attention energy disclosure issue related to modeled heat flux errors interpretation. Below we address the specific remarks of the reviewer.

Sections 1–4

Reviewer: *"The abstract and sections 1 (Introduction), 2 (Lake models), 3 (Observations) and 4 (Experimental set-up) are in general clear and fit for the purpose, except that I am curious as to what precipitation data are used for (p3998, line 28)?"*

Authors: We added the following sentence in the manuscript: "Note that precipitation data are required by some models to compute the mass budget of a lake and the source or sink of heat at the water surface."

Section 5.1

R: *"The presentation and discussion of results begins in section 5.1.1. The division of the timeseries into two periods should be more clearly identified with the shorthand labels of "summer" and "autumn" which are also used later."*

A: We adopted "summer" and "autumn" in most places instead of "period 1" and "period 2".

R: *The data in figure 2 are difficult to distinguish, and perhaps could be expanded in the supplementary material.*

A: We divided the complete time series into two panels instead of three (see new Fig.2) in order to increase the readability.

R: *Please provide a reference for the observations at midlatitude lakes during the warm period (p4008, line 1).*

A: We adopted the sentence as follows "This indicates that the mean bottom heat flux in these models was directed from lake to sediments in the reference run ultimately cooling the surface temperature, which is consistent with water-sediment temperature difference typically observed at midlatitude lakes during the warm period." Thus we point at widely known fact of temperature stratification during summer in midlatitude lakes, that we feel needs no special reference.

R: *There is an unsatisfactory overlap in the presentation of data between sections 5.1.1 and 5.1.2. The data for the FLake model with and without the bottom-sediment effect are presented in figure 3 (and subsequent graphs of this type), but not those data for the other two models for which deactivation of the bottom-sediment effect is tested in section 5.1.2. The presentation style should be unified to show all three models in this way, or none. In my opinion the inclusion of data from both FLake runs in the figures is a useful addition and I would prefer to see all three models thus displayed. Perhaps this could be achieved e.g. by grouping results for each model together and shading the deactivated-sediment results.*

A: As the discussion of the role of bottom sediments is presented in Section 5.1.2, we removed the data of FLake run with deactivated sediments from Fig.2 and Fig.3. Now they are present only on Fig.6 and Table 3.

R: *Minor related points are that the graph labels of ”FLake active sediments” and ”FLake passive sediments” do not seem to be fully explained anywhere, and change to other terminology in figure 6.*

A: Graph labels ”FLake active sediments” and ”FLake passive sediments” are now replaced by “FLake” and “Flake-no-sed”, respectively, and the suffix “no-sed” is now used also for other models, which improve consistency.

R: *Also that the entry in table 3 for FLake, period 1 shows 1.13 but the central panel of figure 3 shows 1.12, and if I am right in assuming that these numbers should be the same then this needs to be checked.*

A: The numbers have been checked and are now the same.

R: *Statistical quantities are defined in the last paragraph of p4005 and I assume that the bias (referred to here as DM) follows the convention of being modelled variable minus observed variable.*

A: DM = mean modelled variable minus mean observed variable. The definition has been clarified in the text (see Section 5.1.1).

R: *However, section 5.1.2 could do with some sort of summing up at the end indicating exactly the conclusions that are drawn from this analysis. Are the effects regime dependent? Model dependent? Is Flake less sensitive to the deactivation of the bottom sediment than the other models, and if so can the authors speculate as to why?*

A: The main results have been added at the end of this Section: “As a conclusion, the results can be summarized as follows: (i) in summer bottom sediments impose minor influence on surface temperature due to stable stratification (the sensitivity of surface temperature to the activation of the sediments routine is about -0.1°C), while (ii) in autumn, the effect of sediments is larger in most models due to complete convective mixing (the surface temperature sensitivity is about 0.2-0.4°C; (iii) the error of summer-autumn surface temperature difference for models lacking explicit sediment inclusion can be explained by 50-100 % through the absence of such a sediment parameterization.”

R: *“In section 5.1.3, figure 4 is useful in interpreting the data shown in figure 5. Attention should be drawn to figure 4 earlier on in this section.”*

A: Figure 4 is now referenced before figure 5 at the beginning of Section 5.1.3.

Section 5.2

R: *At the bottom of p4012, the divergence of the LAKE model from other models in late autumn is attributed to the large thickness (10m) of the soil layer underneath this model. This begs the question as to how water sediment heat exchange is parametrized in this model. Is heat exchanged with the entire mass of sediment, or only through the interface by some parametrization of a diffusive process? If the latter, then the presence of a very deep soil layer ought not to significantly alter results compared to having a layer of only moderate depth. If the former, then this would seem to be a sub-optimal feature of the model which should be highlighted.*

A: The authors agree that sediment/rock depth considered by a model is not the only factor contributing to bottom temperature evolution. LAKE model explicitly considers the diffusion of heat in 10 m sediments, but CLM4-LISSS considers it as well in deeper layer (40 m beneath the bottom). However, CLM4-LISSS does not provide positive bottom temperature bias in the late autumn as it is the case for the LAKE model. We can speculate on the influence of different initial temperature profiles used in CLM4-LISSS and LAKE, which could affect autumn bottom temperatures, and also different properties (specific heat, heat conductance etc.) of sediments in these models, but this shifts the discussion from the main focus of the paper. Hence, we decided to exclude the short discussion on LAKE's bottom temperature warm bias in late autumn from the paper, as this bias is not the most prominent contrast between models shown at Fig.6.

Section 5.3

R: *Figures 7 and 8 appear to be identical and I assume that this is a mistake (the alternative being that the lake models are all perfect, which is unlikely!). The discussion of these results in the text would suggest that the proper data are similar, or that the authors have not noticed that the graphs are identical. However, if the correct form of one figure is actually missing, the assessment of the results by the reader is not possible.*

A: Indeed that was a technical mistake (identity of Figures 7 & 8), that has been fixed in the new version.

R: *The possible underestimation of fluxes by the eddy-covariance measurements is discussed on p4016 as an explanation of the positive bias in the turbulent-flux models. While this may be a contributing factor, other land-based studies comparing surface-flux schemes with such measurements do not describe such a clear systematic bias (the most recent example I can find is Flerchinger et al., J.Hydromet. 2012 volume 13). The Nordbo et al. (2011) reference which is cited indicates that the low-frequency deficit is possibly caused by a lake-forest internal boundary layer and may be minimised in wind directions of maximal open-water fetch, however the present study does not observe any substantial reduction of bias by such a fetch restriction (p4014, lines 14–17). Nordbo et al. (2011, section 2.7) also indicate that "tube attenuation" in their gas sampling system is the dominating cause of LH-flux underestimation, whereas the LI-7500 in the present study is an open-path gas analyzer which does not suffer from this effect. Is it not also the case that Nordbo et al. (2011) obtain their energy balance residual after correcting their flux measurements?*

A: Authors agree that there is no general agreement on heat imbalance problem in eddy covariance measurements. An overview of papers devoted to this problem has been published and added to the references list (Eddy covariance, 2012) . Many of these articles indeed report on heat imbalance in case of heterogeneous surfaces. We have conducted a similar lake model intercomparison experiment for Valkea-Kotinen Lake (manuscript in prep.) that shows similar positive biases for the modeled sensible and latent heat fluxes in respect to eddy covariance data. These results are generally consistent with heat balance residual value presented in Nordbo et al. (2011), a paper that specifically focused on that topic. Moreover, these biases remain significantly large when calculating them only for open-fetch wind directions.

R: *There are ways of estimating or correcting eddy-covariance flux underestimation due to the spectral cut-off. It would be useful if the likely mean error in the covariance measurements could be estimated independently, to see if this is similar to the observed bias, or to the 20–30 Wm^{-2} underestimate proposed on p4016, line 22. Without this, the superiority of the flux models over the flux measurements is debatable. It would also be helpful to estimate the random error in the flux measurements for comparison with the RMSE values presented.*

A: We added the following details to the "Observations" section, addressing the spectral cut-off correction *inter alia*:

- "The eddy covariance flux data have undergone the standard transformations and corrections that are generally recommended for this kind of measurements, i.e., double rotation of the coordinate system, buoyancy and cross-wind correction of the sensible heat flux, density correction of the latent heat flux and a simplified correction for spectral losses to which sensor separation between the sonic and the IR hygrometer is the most important contribution.” As to independent estimate of mean error of eddy covariance measurements, we refer to EC-sensors intercomparison presented in (Mauder et al., 2006).

- The following text is added to the paper: “Another possible source for systematic underestimation of real fluxes by the EC technique is cause by sensor-specific errors (Mauder et al., 2006). It is shown by these authors that DM ("the bias" in the paper) of total heat flux for the USA-1/LI-7500 system used at the Lindenberg Observatory in respect to a reference device was -4 W/m^2. The total heat fluxes for other devices taking part in this study deviated on average by 10-30 W/m^2 from the reference one. This range of uncertainty due exclusively to sensors design suggests that the heat imbalance in our study can be significantly affected by this factor as well.” Shortly, this means EC measurements may provide systematic bias in respect to true fluxes due to both EC devices design and flux calculation procedure.
- As to estimates of RMSEs of eddy covariance measurements, we included the following text in our manuscript “It was demonstrated that RMSE ("RMSD" in (Mauder et al., 2006)) of sensible and latent heat fluxes for the USA-1/LI-7500 system of Lindenberg Observatory used at Kossenblatter See, in respect to a reference one (CSAT3/LI-7500) were 15 W/m^2 and 24.4 W/m^, respectively. These values are close to model errors estimates (Fig.~7 and 8) and thus prevents us from deriving any further definite conclusions on flux schemes quality based on comparison with eddy covariance data in terms of RMSE.”

R: *On p4016 the discussion of the term R representing advective and diffusive contributions to the heat budget is slightly speculative. The possible existence of seasonal circulations and their potential for providing a net heat flux is discounted, although one could imagine possible scenarios where this occurs, particularly for a marginal location like that of the observations. It is also unclear what the flux rate of water through this lake is. It may not have a major inlet, but it has an outlet. Is this gauged, or is there a gauge further downstream that may be used to estimate the outflow? What is the ratio of the lake area to the lake catchment area? Either of these may help provide a quantitative estimate of advective strength.*

A: In last Section we included a simple estimate of the magnitude of heat advection strength due to inlets based on lake surface area, lake catchment area and climatological amount of precipitation. This brought us to an upper estimate of 9 W/m^2 in terms of surface flux that is 3 times less than a heat balance residual. The seasonal averaged effect of advection by inlets is likely to be less than this upper estimate. As to the role of internal circulations we do not speculate on this issue since a more solid basis for any conclusions would require the use of 3D lake models which we leave for future research.

Technical corrections

We adopted all of them.