Author response to Referee #1 comments for Manuscript #HESS-2020-222, "Variability in epilimnion depth estimations in lakes" by Harriet L. Wilson, Ana I. Ayala, Ian D. Jones, Alec Rolston, Don Pierson, Elvira de Eyto, Hans-Peter Grossart, Marie-Elodie Perga, R. Iestyn Woolway, Eleanor Jennings.

Thank you for your valuable comments which have enabled us to improve our study. Please find below our responses to each comment respectively. Updated text is in italics.

1. Referee comment: The article evaluates the epilimnion depth estimate in high-frequency data made by four different methods, including the effect of defining different thresholds that these methods require the user to choose. It also made this estimate using a hydrodynamic physical model. The aim of the study was to highlight the variability of the epilimnion depth estimates and how this variability could impact inferences about lake processes. This study draws attention to the need for researchers to unify their consensus on this topic, allowing comparisons between the results of different studies.

This is a very important study that addresses a hot topic in limnology and oceanography. With the increase in studies evaluating different bodies of water and the ease of obtaining increasingly sensitive measuring equipment, with the ability to perform and store an increasing number of measurements, we have entered the era of Big Data. With the increasing availability of data, the need for models capable of extracting the correct information from them also increases. The correct adjustment of these models depends on studies of this type.

Author response: Thank you for these positive comments. We greatly appreciate the enthusiasm for the study as a topic of current relevance.

2. Referee comment: I believe that the result obtained and discussed in the article mixed the estimates made in clearly stratified water columns, with estimates made in weakly stratified water columns with estimates made in water columns with the presence of multiple stratifications. Therefore, the described variability is not only due to the distinction between methods and limits, but also to the application of the methods under different conditions, and this should be clearer.

Author response: This is an important point by the reviewer, and we agree that both 1) the presence of multiple stratifications and 2) weakly stratified profiles, contribute to the large variability found between epilimnion depth estimates. In fact, it was *particularly* when water column profiles did not conform to the idealised three-layered structure, that we found the greatest divergence between epilimnion depth methods. However, we would argue it is not simply a matter of filtering out weakly stratified or multiple stratification profiles, since this would result in large data gaps, including in peak summer, and is limited by the same issues of subjectivity as the definition of the epilimnion depth itself. Instead, there is a need for users to acknowledge both the systematic differences between methods, and the application of the methods under different conditions, in respect to their specific purpose. We have addressed this point in the revised manuscript (please see Comment #3 for relevant edits to text).

3. Referee comment: P. 8 L. 281-283: I think that authors should make a clearer distinction between primary and secondary thermocline. For example, the method described by Read et al (2011) allows us to estimate the two thermoclines and, therefore, is more sensitive than the other methods, which do not consider this possibility. The comparison between the methods must consider the presence of these superficial micro-stratifications. For example, in the graphs b) of figure 4, there is clearly the presence of these microstratifications that are hampering some methods of identifying the main thermocline. In other words, it is not fair to fit a model that expects only one stratification with profiles that show various stratifications. It is obvious that the estimate will not be satisfactory. It is extremely necessary to make a pre-filter, removing the superficial layers of values before the estimate is made. This method was applied by Pujoni et al. (2019).

Author response: See the response to Comment #2 above and Comment #4 which are related. We do not think it is appropriate in this study to make a defined distinction between primary and secondary thermoclines, or filter/smooth out these stratifications for the following reasons; 1) determination of primary/secondary thermoclines is not objective and the selection of profiles requiring smoothing would demand further arbitrary thresholds (instead we argue that this topic deserves a separate study, investigating the full implications of different approaches), 2) in this study we aimed to estimate variability in epilimnion depth estimates between *common and existing* methods (rather than introduce new methods), where issues related to microstratification are a result/discussion point, 3) smoothing of these microstratifications may not be suitable for some applications, particularly where it is assumed that the epilimnion is isothermal and well-mixed. Nevertheless, we do fundamentally agree with the reviewer that this is an important issue that should be clearer in the text, therefore we propose editing the discussion to emphasise these points (see below). We also now cite the work of Pujoni et al., (2019) and Read et al., (2011) in this context.

L.388: '*The concept of the epilimnion, and more widely, the three-layered structure of a stratified lake, is fundamental in limnology. Yet, despite the ubiquity of the term, there is no objective or generic approach for defining the epilimnion and a diverse number of approaches prevail in the literature. In a comprehensive analysis of high-frequency, multi-year data from two lakes, this study has highlighted the extent to which common water temperature profile based epilimnion depth estimates differ. The level of variability in epilimnion depth estimates calculated using common methods and threshold values, was exceedingly high. This result calls into question the practice of arbitrary method selection and comparing findings between studies which use different methods or even just different thresholds. The magnitude of variability also casts ambiguity on the calculation of key biogeochemical and ecological processes in a lake that rest on the assumption that the layers of a lake are well defined, including calculations of metabolic rates, and oxygen fluxes (e.g. Coloso et al., 2008, Foley et al., 2012, Obrador et al., 2014, Winslow et al., 2016).*

In an idealised stratified profile, the epilimnion is portrayed as near-uniform in water temperature or density and clearly delineated from a well-defined metalimnion. However, many measured profiles, at least within this study, did not conform to this idealised three-layered structure. Instead the water columns were often more complex, including multiple pycnoclines and near-surface micro-stratification layers, or the boundaries between the epilimnion/metalimnion were blurred. One approach to this issue is to filter out appropriate water column profiles or apply functions that coerce the profile into the expected structure (Read et al., 2011, Pujoni et al., 2019, Gray et al., 2020). Filters, additional conditions or smoothing functions, however, may suffer from many of the same challenges as the estimation of the epilimnion depth, since they attempt to discretise data based on arbitrary criteria (Kraemer et al., 2020). For example, our analysis of temporally high resolution time series data emphasised that rather than jumping from states, such as stratified or isothermal, changes in the water

column occurred over an evolving continuum and often fluctuated between states. Similarly, the distinction between additional layers, such as the primary or secondary pycnocline, is fraught with the same issues of arbitrariness as discussed (Read et al., 2011). This study demonstrates that when epilimnion depth estimation methods, which are theorised for a three-layered water column, are applied to non-conforming water columns, they diverge widely on the location of the epilimnion depth, and at times, may not even be underpinned by the same theoretical assumptions. Since none of these methods can be considered the 'true' definition of the epilimnion depth, it is necessary to understand the degree to which methods differ. Improved understanding of their systematic differences will facilitate the use of methods that appropriately capture different processes, such as, air-water exchanges, thermocline entrainment or suspension of materials. Due to the realised complexities of observed and aggregated profile data, we may benefit from new approaches to water column discretisation that incorporate the vast proportion of profiles which do not conform neatly to the three-layered paradigm.'

4. Referee comment: P. 8 L. 288-289: In this same line, we must discuss and define a threshold of what we call "homogeneous water column". I don't think it makes sense to compare the methods using profiles with low stability of the water column. If we no longer have a clear stratification, the methods should not be applied, as they will look for a thermocline that does not exist. I may be wrong, but the water column in the graphs c) in figure 4 is homogeneous and should not be subjected to comparison with these models.

Author response: This is another great point by the reviewer, and we agree that the stability of the water column has an influence on the results. Inherently, however, there are conditions within each of the methods, relating to the degree of stratification required to estimate the epilimnion depth. Method 1 has the precondition that the range in water density must be greater than the threshold value, else the epilimnion depth is assigned to the maximum lake depth. Similarly, Methods 2 and 3 have the precondition that the water density gradient must be greater than the threshold value. Finally, Method 4, the rLakeAnalyzer, is slightly different as it initially filters out profiles based on a 1°C water column range and will then identify the maximum density gradient regardless of the threshold value. As discussed inComment #3, a further stability-based condition could be introduced, but would suffer from being an arbitrary threshold to which the rest of the results would become dependent. Again, it is the fact that studies are currently using different approaches with different inherent stability thresholds that can contribute to the confusion caused when comparing studies, and this is one of the key points we are raising in this manuscript.

Note, though, that we did extensively investigate the use of pre-filters, including top-bottom density differences, water column total density gradient and Schmidt stability values. However, echoing the presented findings of the epilimnion depth analysis, we found different methods and thresholds largely altered the period that was deemed stratified. This analysis could not be justly presented in this study, without overly complicating the manuscript. In addition, they resulted in the removal of large amounts of data, even within peak summer, which is not suitable for analysis of mixing events for example. Finally, presenting the full time series results demonstrated the perils of using temporal means of epilimnion depth. For example, for calculating the summer mean epilimnion depth, the un-filtered mean will be influenced by periods of very deep epilimnion depth estimates (i.e. when the water column is nearly isothermal), while the filtered mean would not be representative of conditions during the full summer, but rather a subset of the stratified profiles.

5. Referee comment: P. 9 L. 350: Why did the authors use the range to estimate variability? The range is sensitive to outliers. Why not use standard deviation, which is a more robust estimate of variation?

Author response: The reviewer has highlighted an important point and in the revised document we will use inter-quartile range which more robust to outliers than range. To address this we update the methods, results (Table 2 and Figure 7), although the findings are very similar.

Table 2. Summary of statistics for each method, showing the mean (m), minimum (i.e. shallowest estimate) (m), maximum (i.e. deepest estimate) (m) and the interquartile range (m) of the April-October epilimnion depth estimates (summarised from the results shown in Fig.6a), and the mean Pearson's correlation coefficient (r) for each method, representing the mean correlation for all possible combinations between threshold values., for Lough Feeagh and Lake Erken.

Method	Lough Feeagh					Lake Eken				
	Mean (m)	Min (m)	Max (m)	IQR (m)	r	Mean (m)	Min (m)	Max (m)	IQR (m)	r
M1	-19.0	-4.6	-25.4	7.3	0.77	-10.0	-7.8	-11.2	1.3	0.92
M2	-35.9	-19.7	-41.4	6.5	0.48	-11.3	-8.4	-12.9	1.7	0.78
M3	-36.5	-22.4	-41.5	6.1	0.49	-11.8	-10.0	-13.0	1.2	0.82
M4	-21.1	-19.7	-21.5	0.3	1.00	-11.9	-10.1	-11.3	0.5	0.99

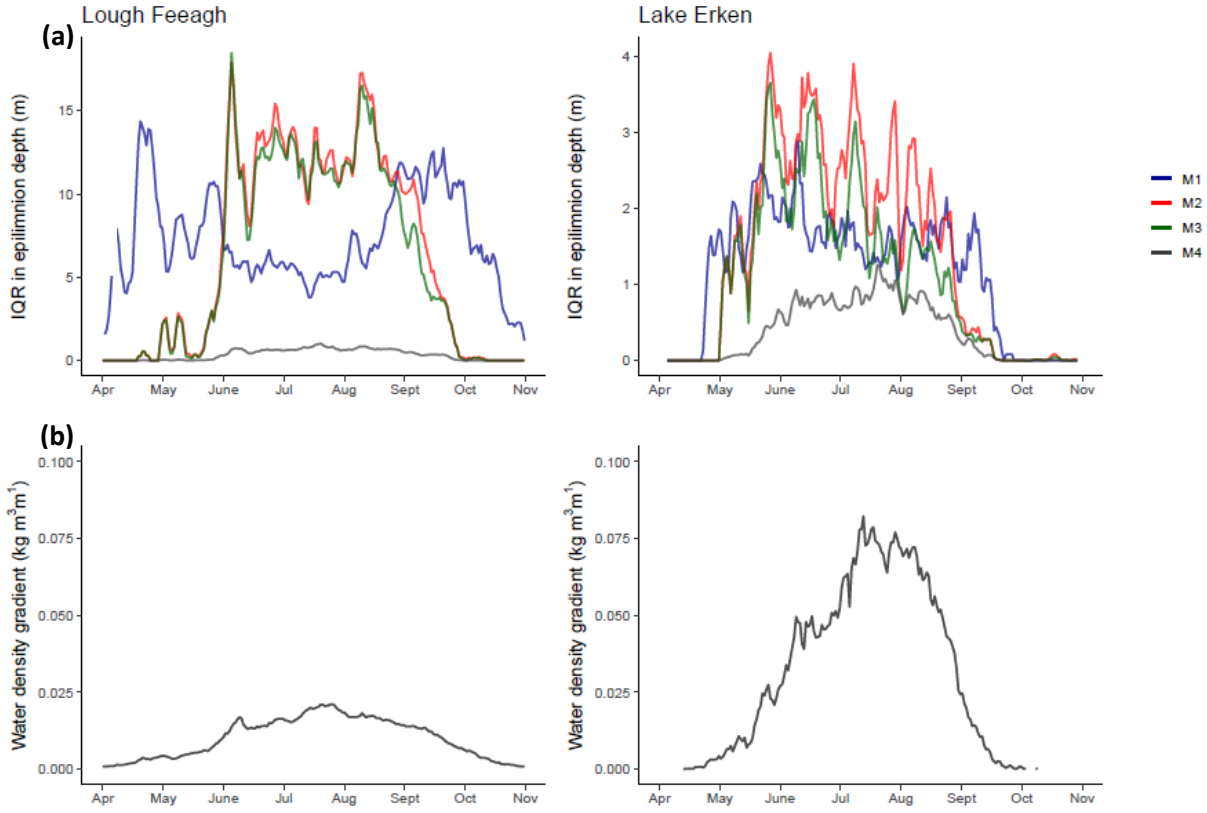

Figure 7. Inter-quartile range between the shallowest and deepest estimate for each method calculated from long-term daily mean epilimnion depth estimates for each Julian day, where a large range suggests high threshold sensitivity and a small range suggests low sensitivity **(a)**, and long-term daily mean water density gradient, calculated based on the surface and maximum measured depths **(b)**, for Lough Feeagh and Lake Erken.

L. 225-227: *'We also summarised these statistics for each method, showing the mean, minimum (shallowest), maximum (deepest) and interquartile range for each method, to demonstrate differences between methods. A large interquartile range in epilimnion depth estimates, indicated high sensitivity to the threshold value.'*

L. 317 – 323: *'For both lakes, the interquartile range in the mean Apr-Oct epilimnion depth estimates for each method was very high for M2, M1 and M3, indicating high threshold sensitivity in these methods. Method M4 had a substantially lower interquartile range than all other methods and a very high mean Pearson's correlation coefficient, indicating that both the mean value and the temporal pattern of the epilimnion depth were only weakly influenced by the threshold value. In both lakes, methods M2 and M1, where the epilimnion depth was defined from the surface downwards, had a higher interquartile range in estimates calculated with different threshold values, compared to methods M3 and M4, where the epilimnion was defined from the pycnocline upwards.'*

L.348 – 366: '*For all methods, threshold sensitivity fluctuated seasonally, although varied in pattern (Fig. 7). Threshold sensitivity was shown by the interquartile range between the shallowest and deepest epilimnion depth estimates calculated for all threshold values. In Lough Feeagh, M1 had a smaller range in epilimnion depth estimates during the peak summer months of June, July and August, compared with months when the onset and overturn of stratification commonly occurred. During periods of transient stratification, the stability of the water column was often low but frequent changes in the near-surface water density, induced large differences between estimates calculated using small thresholds compared with large threshold values. In contrast, methods M2 and M3 had the highest range in estimates occurring during the peak summer months. Even during peak summer in Lough Feeagh, gradients in the water column were relatively small (Fig. 7b), which resulted in a very large range between the smallest threshold values which found a near-surface epilimnion depth, and the largest thresholds that often found no epilimnion depth at all, therefore defaulting to the deepest depth. In Lake Erken, the water density gradients were typically much larger, and methods M1, M2 and M3 all peaked during May and June, when gradients in the water column were typically increasing but prone to fluctuations. For both lakes, M2 had typically a higher threshold interquartile range than M3 during peak summer and the overturn period, which was related to the common development of a secondary pycnocline. M4 produced much lower interquartile ranges in the epilimnion depth throughout the year, since as long as the 'mixed.cutoff' filter was met, the epilimnion depth was defaulted to the pycnocline if the threshold was not exceeded, thus largely reducing the ability for large differences to occur. The interquartile range in epilimnion depth estimates for M4 was highest during the peak summer months, which was when the epilimnion depth was typically shallowest and more frequently defined by the threshold value rather than defaulting to the pycnocline.'*

6. Referee comment: P. 10 L. 379-382: I would suggest showing some graphs of density profiles with the estimated depths of the methods so that it would be easy to see why there were such differences and whether one method made a "better" estimate than the other. I would suggest that the authors discuss a little about the visual assessment of profiles. Should we rely on this visual assessment to try to "correct" the biased estimates made by the models?

Author response: Visual assessment of profiles is certainly very helpful and is also commonly practised in limnology. We think that with 5 tables and 7 figures it may be excessive to add additional figures. Our intention with Figure 4 was to demonstrate to the user all methods/thresholds on three distinctly different profiles, however, visual assessment is not part of our result analysis. The focus for this study is for using high-frequency data and multi-lake analysis, and therefore the goal is to find methods that can be used systematically without being tailored to specific lakes.

We are interested however in visual assessment of epilimnion depth and we have conducted a survey investigating where limnologists visually identify the epilimnion depth using profiles from anonymous lakes. This is something we would like to publish at a later date as a short discussion paper, but is not appropriate for automated analysis of high frequency data.