Response to referee comment Referee #3

We appreciate and would like to thank Referee #3 for taking the time and effort to read our manuscript and expressing the generally positive impression of our work. We will improve our manuscript based on reviewer's helpful comments. Our point-to-point response are below (comment of the referee in black, our response in purple).

Point-to-point response

General evaluation and major comments

R3-1: Overall, I think the manuscript is well structured and the methodology is well prepared. The presented tables and figures support the findings that were presented in the text and provide good insights in the functionality of the presented algorithm. While I think the overall quality of the manuscript in its present form is already good, I see some crucial points that require clarification. Other less significant points should also be improved to improve the quality and readability of the manuscript. I will outline my major concerns in the following and will address smaller issues in a line-by-line notation in the following section.

Response: Thanks for reviewer's comment. We will revise the manuscript based on the below comments raised by reviewer.

1. *Synthetic study design:*

R3-2: The observation data were generated with the same model structure that was later on investigated in the case study. Which means that "observed" variables (in this case velocity and temperature) were calculated with the exact same set of equations that in the following calculated the simulated variables. I think this property of the observation data could potentially impair the entire analysis and favour the simultaneous calibration of velocity and temperature.

As the authors outline in section 3.5 of the manuscript, multi-variable calibration problems are often affected by trade-offs between the variables for which performance metrics should be optimized (minimized). One reason for that is that models are simplification of the represented reality and parameters can affect multiple processes, sometimes in the opposite directions. Thus a change of a parameter value into one direction could improve the performance of one metric, while deteriorating the performance of another metric at the same time. This is different in the synthetic example where in fact the reality and the model are the same thing. Thus the "observed" time series of velocity and temperature perfectly agree with the model simplifications and assumptions.

Thus, in such a case providing both, temperature and velocity, to the search algorithm should better constrain the parameter response surface than only providing one of the two variables. If this hypothesis is sound, then the presented results would by affected by this effect. In this case only a real case scenario could provide an honest comparison of the three cases.

Response: The reviewer questioned the rationality of using data generated from the model as observations. We fully agree that using real observation data for such investigation is better. However, real observations for velocity are not available for this study, and are usually difficult to

find. Due to scarce availability of observed velocity data, the use of both temperature and velocity for calibration of 3D lake hydrodynamic models is also rare both in practice and research. However, understanding of the importance of velocity data in calibrating such models is important and could be possible by generating synthetic observed data from lake hydrodynamic model simulations. These underlying hydrodynamic models are physically-based and precise. Hence, we believe that generating synthetic observation data from such models is reasonable. We agree with the reviewer that generating synthetic data from conceptual environmental simulation models may not be reasonable though (e.g., conceptual hydrologic models that have intrinsic structural errors and uncertain conceptual parameters). Of course, even in the case of 3D lake hydrodynamic models, synthetic model scenarios cannot be considered the same as real-world situations, but they are close representations of real-world situations. We think that our analysis on model-based synthetically generated observed data can provide at least some implications for the investigation people could do with real-world data. Our research implication of the use of velocity observation is also in line with the study of Baracchini et al. (2020), where they also suggest having both temperature and current velocity for complete system calibration.

Baracchini, T., Hummel, S., Verlaan, M., Cimatoribus, A., Wüest, A., & Bouffard, D. (2020). An automated calibration framework and open source tools for 3D lake hydrodynamic models. Environmental Modelling & Software, 134, 104787.

The reviewer also highlighted the discussion of trade-offs in section 3.5. We need to clarify that section 3.5 discusses other possible applications for which our new method, DYNO, could be used. But we don't think the multi-variable hydrodynamic calibration problem we are dealing with is a trade-off problem. As we have discussed in the introduction (lines 113-122 of the original manuscript), it is not apparent that there is usually a trade-off between the fit of multiple variables / constituents. This might be true for conceptual models, e.g., hydrologic models, where equations are experimental, and there is lots of uncertainty embedded in parameterization. In such models, trade-offs typically exist between different sub-objectives (e.g., using different performance metrics). But hydrodynamic models are physically based. Moreover, lake water velocity is important for determining distribution of water temperature. Hence, we believe that simulating velocity more accurately in hydrodynamic models, is expected to improve accuracy of water temperature simulation and distribution. Consequently, a trade-off may not exist between the fits of temperature and velocity. We agree that investigation on real data might be more convincing in ensuring the above hypothesis as well, and this could be part a future study where observed data for both temperature and velocity is available. We will revise our statements in the manuscript to be more cautious about implications for real-world practice based on modeling results.

2. Number of iterations of the search algorithm

R3-3: This comment is somehow related to the previous one. In theory it should be possible for the search algorithm to identify the "true" parameter set that was used to generate the synthetic observation data. This global minimum is present on the parameter response surfaces of all three calibration scenarios and has a value of 0 independent of the used metric (DYNO or the single performance metrics). Thus, when not ending up in a local minimum all three cases should converge towards this global minimum. As outlined above, I simply think that the Cal-Both scenario does this quicker due to the given reasons.

In the presented results all "best" solutions did not find the global minimum. As far as I got it right from the text, each experiment involved 8 iterations with 24 parallel evaluations in each iteration step. Given a computation time of 5 hours per simulation run (according to the text and thus 5 hours per parallel iteration) one experiment takes 1.6 days. I am wondering if experiments with larger numbers of iterations were performed (that are maybe just not shown). I would be interested how the convergence of the three calibration experiments develops with larger number of iterations.

Response: The reviewer brings up an interesting aspect to discuss. In theory, the global minimum has a value of 0 for all three calibration scenarios. However, it might not be correct that having zero temperature error at these observation locations would mean zero velocity error at these observation locations. This is because only the observation at part of the simulation space is used for calibration (not the observation data at each grid and each time step of the simulation space are used to calculate the temperature error). So the temperature error may be 0 at these observation locations, while the temperature error is not 0 at other locations where observation is not used in calibration. In this case, getting a temperature error 0 at observation locations cannot guarantee the velocity error is 0. Hence it is possible that the optimization ends up in a local minimum and does not converge towards the global minimum (both temperature and velocity error are zero) in cases calibrating to only temperature or velocity. In optimization search, the algorithm might not even reach the solution with zero temperature or velocity error when calibrating to only one variable. This is because the optimization only focused on one variable (temperature or velocity) and ignored the calibration on another. For example, if only calibrated to temperature, the velocity error might still be large. In this case, the algorithm is less likely to find the solution with temperature error 0 while the velocity error is large. This is exactly what Figure 5 (a) in the original manuscript showed (We copied Figure 5 (a) below). In Figure 5 (a), when calibrating to temperature only, the best solution found after each iteration is converged at places where the velocity error is large, and temperature error is not reduced to 0. In contrast, the optimization search considered both temperature and velocity error when calibrating to temperature and velocity. Hence the best solution found after each iteration is improved in both temperature error and velocity error.

The reviewer asks for experiments with a large number of iterations. We did not increase the number of iterations because, 1) we found the optimization search almost converged at the last few iterations (As shown in the calibration progress plot Figure R1 below) and 2) our problem is computationally very expensive; one single simulation takes about 5 hours; 24 simultaneous simulations take about 8-12 hours because of limited memory resources. It takes more than 64 hours to conduct one calibration experiment, and we have three scenarios, and each scenario is repeated in three trials. This is a huge amount of computing. Meanwhile, in real practice, users are also unwilling to wait too long for the calibration process. That is why we compared the accuracy of the solutions within a limited computing budget.

In the revised manuscript, we will add the above discussion into the manuscript, which we think is interesting.

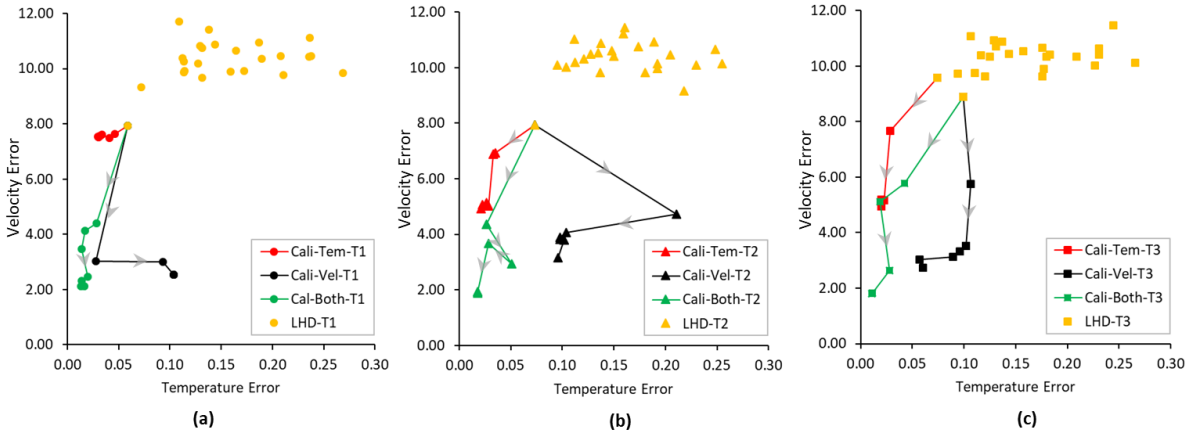

Figure 5 (Copied from the original manuscript). Calibration progress plot of the best solution found (in term of objective function value) during optimization search by PODS when calibrating to temperature only (Cali-Tem), calibrating to velocity only (Cali-Vel), and calibrating to both temperature and velocity (Cali-Both). Three random trials (i.e., T1, T2, and T3) are plotted in (a), (b), and (c). Lower velocity and temperature error are better. The yellow makers are evaluation point in initial experiment design using Latin Hypercube Design (LHD). Besides solutions in LHD, only the best solution in each of the optimization iterations are plotted (i.e., makers lined with lines). The line links the best previous solution in one iteration to the best solution in next iteration. The arrow indicates the direction from the previous solution to the next solution.

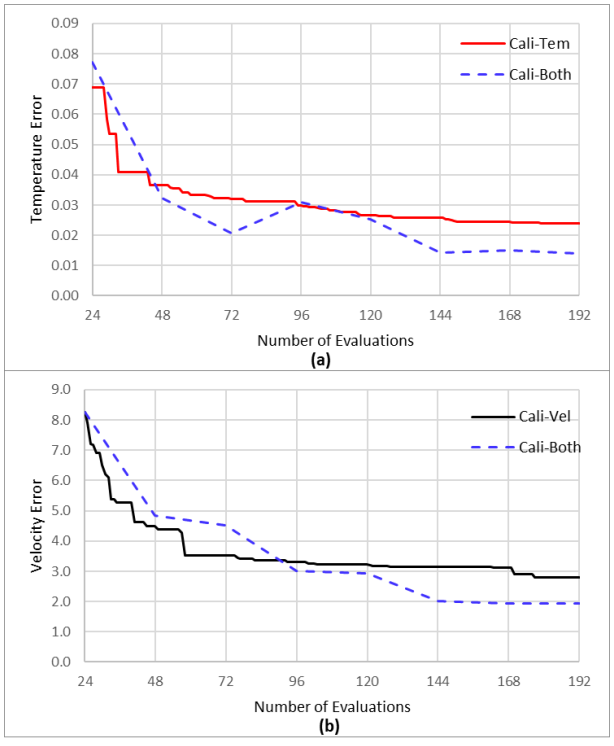

Figure R1. Calibration progress plot of PODS in Cali-Tem, Cali-Vel, and Cali-Both. The best solution found so far in average of three trials is plotted with the number of evaluations.

3. Consistency in the nomenclature

R3-4: Overall I think the outline of DYNO and the explanation of variables is well done. Yet, I found some inconsistencies in the nomenclature and some mathematical definitions that must be improved. I will address these (at least the ones that I found) in the line-by-line comments.

Response: Thanks for the reviewer's comments. We will carefully revise the nomenclature and mathematical definitions. We also have a detailed response to the line-by-line comment below.

4. English language

R3-5: Overall I think the manuscript is well written. In some sections of the manuscript I found that the same errors repeat in every second line (e.g. singular/plural, missing articles). Although I am not a native speaker myself I had the impression that the manuscript could require some proof reading. Some sentences I had to read over and over again, but they still do not make sense to me. I addressed them as well in the line-by-line comments.

Response: Thanks for the reviewer's comments. We have carefully looked at the reviewer's comments about these language issues. We will revise them as the reviewer suggested in the revised manuscript.

5. Non-color blind color theme in figures

R3-6: Although this does not affect me, I would advise to change the color theme in the figures and refrain to use red and green at the same time (8% of males are affected by red-green color blindness).

Response: Thanks for the reviewer's comments. We agree with the reviewer's suggestion, which is very important. We will change the red and green to red and blue colors for all related figures.

Line-by-line comments

p.1 L13, p.4 L123 and further: Please use a consistent naming for Dynamically Normalized Objective function. Either all first letters caps or none. It differs throughout the text. Also only use the acronym after it was first introduced in the text.

Response: We will revise the manuscript with a consistent naming for Dynamically Normalized Objective Function and used acronym after the first introduction in the text.

p.1 L21 - 22: Please rephrase this sentence. Further, I think the statement is not universally true how it is formulated, that calibrating only one variable does not improve the calibration of another variable.

Response: Our statement is not to say that calibrating one variable does not "improve" the calibration of another. We say calibrating one variable cannot "guarantee" the fit of another. Our results support our statement. When calibrating to temperature only, the velocity error is large, and vice versa. It is true in our case. Of course, it might not be universally true for other cases. We will rephrase the sentence to be: "The result indicates that DYNO can balance the calibration in terms of water temperature and velocity and that calibrating to only one variable (e.g., temperature or velocity) cannot guarantee the goodness-of-fit of another variable (e.g., velocity or temperature) *in our case*."

Eq. 2, Table 1: Consistent notation of *Sim* and *Obs* necessary. Either capital first letter or lower case.

Response: We will capitalize the notation of Sim and Obs in Table 1 and throughout the manuscript.

Table 1: Please rephrase the sentence: "The set of variables the observation data of which is used in calibration"

Response: We will rephrase the sentence to be: "The set of variables whose observation data is used in calibration."

Table 1: I would suggest to add the vector $\{x_1, ..., x_d\}$ to the definition of $\boldsymbol{X}$ as e.g. Table 2 refers to them.

Response: We will add following text into the definition of X in Table 1: "$\boldsymbol{X} = (x_1, x_2, \ldots, x_d)$."

Table1 and throughout the text: I would suggest to use the nomenclature $Sim^{k}_{j,t}$ (same for *Obs*) in the definition and throughout the text to indicate the time dependency. This nomenclature is in fact introduced later in the manuscript. Also the definition refers to the time *t* which is not indicated in the variable.

Response: We will add the definition of $Sim^{k}_{t,j}(\boldsymbol{X})$ and $Obs^{k}_{t,j}$ in Table 1. We will bold the definition of $\boldsymbol{Sim}^{k}_{j}(\boldsymbol{X})$ and $\boldsymbol{Obs}^{k}_{j}$ to denote the vector of simulation output and observations data at all time steps ($\boldsymbol{Sim}^{k}_{j}(\boldsymbol{X}) = \left(Sim^{k}_{1,j}(\boldsymbol{X}), \ldots, Sim^{k}_{N,j}(\boldsymbol{X})\right)$ and $\boldsymbol{Obs}^{k}_{j} = (Obs^{k}_{1,j}, \ldots, Obs^{k}_{N,j})$). We keep the definition of $\boldsymbol{Sim}^{k}_{j}(\boldsymbol{X})$ and $\boldsymbol{Obs}^{k}_{j}$ for the simplicity of Eq. (2) (since the goodness of fit function takes the simulation output and observation vector as input).

p.5 L153: ...multiple variables *with* a single objective function.

Response: We will revise the sentence to be: "… formulate the error of multiple variables with a single objective function".

p.5 L159: Please remove \textit{value} as the distribution of NRMSE is not a single value.

Response: We are trying to make two comparisons: one is the highest attainable NRMSE value another is the distribution of the NRMSE value. We will revise the sentence to be: "However, it is still possible that the highest attainable value (or distribution) of NRMSE (or KGE) across the parameter space for one variable maybe be much higher than the highest attainable value (or distribution) of NRMSE (or KGE) of another variable.

p.5 L160: Same as above.

Response: We have resolved this comment in response to the comment above.

p.5 L160: Hence **,** ...

Response: We will revise the sentence to be: "Hence, … ."

p.5 L163: Again consistent naming of DYNO or removing the full name as the acronym was already introduced.

Response: We will remove the full name of DYNO.

p.5 L166: ...all evaluations in $\boldsymbol{\psi}$ found so far?

Response: $\boldsymbol{\psi}$ is the set of evaluations found so far by the optimization. We will revise the sentence to be: "…. Let $\boldsymbol{\psi}$ be the set of evaluations found so far by the optimization, DYNO (as shown in Eq. (3)) normalizes the error of each variable $f_k(\boldsymbol{X})$ with its upper and lower bound, $f_k^{max}$and f_k^{min} of all evaluations in $\boldsymbol{\psi}$"

p.5 L166: The variable ψ was not introduced at this point and is I think not defined in the manuscript.

Response: We will add the definition of $\boldsymbol{\psi}$ in response to the comment above.

p.5 L167: *The* mathematical formulation...

Response: We will revise the sentence to be: "The mathematical formulation ..."

p.6 L172: ...all evaluations ...

Response: We will revise the sentence to be: "… all evaluations …"

p.6 L172 - 174: Please rephrase this sentence. It is in my opinion not clear what is meant.

Response: We will break the sentence to two short sentence: "where $f_k^{max}(\boldsymbol{X})$ and $f_k^{min}(\boldsymbol{X})$ are the maximum and minimum values of $f_k(\boldsymbol{X})$ for all evaluations in . $f_k^{max}(\boldsymbol{X})$ and $f_k^{min}(\boldsymbol{X})$ have to be updated dynamically in each iteration during optimization."

p.6 L180 - 185: The procedure as described is confusing and does not really make sense to me. A calibrated model exists from a previous study. This calibrated model was taken, but different parameters based on "expert guessing" were then used to run the model. So it is not the calibrated model anymore?

Response: We will revise the words "expert guessing". The set of "true" value we use is the manual calibration results from expert. The words "expert guessing" might cause confusion. We will replace it with "manual calibration by experts".

p.6 L186: Replace *a* by *the* as you did not use a parameter set but the parameter set $\mathbf{X}^{\mathrm{R}}$.

Response: We will revise the sentence to be: "the vector of model parameters $\boldsymbol{X}^R$"

p.6 L193: ...the true values of the parameter vector $\mathbf{X}^{\mathrm{R}}$ *are*...

Response: We will revise the sentence to be: "...the true values of the parameter vector $\mathbf{X}^{\mathrm{R}}$ *are*...".

p.6 L194: I would replace *value* with *set*.

Response: We will replace value with set.

p.6 L202 - 203: This sentence is not clear to me.

Response: We will revise the sentence to be: "We have the five sampling locations across the reservoir. The observations data at these five locations are used to calibrate the model parameters."

p.7 L226: coefficient*s*

Response: We will change coefficients to coefficient: "The vertical exchange of horizontal momentum and mass is affected by vertical eddy viscosity and eddy diffusivity coefficient."

p.8 L229: *the* Manning formulation

Response: We will add "the".

p.8 L229: ...which is a parameter tha should also be calibrated.

Response: We will revise the sentence to be: ", which is a parameter that should also be calibrated."

p.8 L232: ...*are* parameterized...

Response: We will revise the sentence to be: "...*are* parameterized..."

p.8 L237: *M*anning coefficient

Response: We will capitalize the word "Manning".

Section 2.5: This section is actually a repetition of L146 – 150

Response: We will remove the repeated information. Section 2.5 gives the detailed calibration formulations. We will revise the first paragraph of section 2.5 to be: "Three scenarios are considered to investigate the impact of model calibration against temperature and/or velocity observations (as discussed in section 2.1). The first two scenarios calibrate to only one variable, and the last scenario calibrates both variables simultaneously. This section give the detailed calibration formulations of these three scenarios."

p.9 L280: Remove *them*

Response: We will remove "them".

p.9 L287: I would suggest to replace *a roughly* by *an approximately*

Response: We will replace "a roughly" by "an approximately".

Caption Table 3: formulation*s*

Response: We will change “formulation” to “formulations”.

p.10 L309: Use the acronym DYNO instead

Response: We will replace “the Dynamically Normalized Objective Function” with “DYNO”.

p.10 L317: Replace *is* with *are*

Response: We will replace “is” with “are”.

Figure 2: The positions of the labels *Yes* and *No* in the stopping criteria are not clear to me.

Response: We will correct the position of “Yes” and “No” in Figure 2.

p.12 L359 - 366: The argumentation in this paragraph could be affected by the affect of synthetic observation data as outlined in the first major remark.

Response: We have responded the first major remark in response to R3-1. The discussion in line 359-366 is stating the results based on physical models. These hydrodynamic models are built based on physics and knowledge human learned in the past hundreds of years. Hydrodynamic models are not like most of hydrology models that with lots of uncertainty. They of course cannot be try the same as real world situation but they are close representations of the real world situation. We think the investigation on models can provide at least some implications for the investigation people could do with real world data. Our research implication of the use of velocity observation is also in line with the study of Baracchini et al (2020), where they also suggest have both temperature and current velocity for a complete system calibration. We will add these discussion after L359-366.

Baracchini, T., Hummel, S., Verlaan, M., Cimatoribus, A., Wüest, A., & Bouffard, D. (2020). An automated calibration framework and open source tools for 3D lake hydrodynamic models. Environmental Modelling & Software, 134, 104787.

p.13 L395: Missing end '.' of sentence?

Response: We will add the missing end “.”.

p.13 L395 - 399: This sentence is confusing and might require rephrasing.

Response: We will rephrase the sentence to be: “Figure 3 illustrates that calibrating to temperature data only (red scatter plot) results in larger velocity errors $\Delta\overrightarrow{Vel}$, relative to velocity errors when calibrating to velocity data only (Cali-Vel scenario, i.e., black scatter plot) or to both velocity and temperature data (Cali-Both scenario, i.e., green scatter plot).”

p.14 L406 - 416: I had the impression that many articles were missing here. This paragraph is hard to read in general and could potentially be revised.

Response: We will revise this paragraph to be: “Figure 4 shows the temperature error of solutions from three different calibration scenarios: Cali-Tem (red time-series), Cali-Vel (black time-series) and Cali-Both scenarios (green time-series). The errors between simulated and observed water temperature at

the surface, middle and bottom layers of two stations (STN. A1 and STN B1) are plotted. In general, the temperature error of the solution in the Cali-Both scenario is generally close to zero °C for all the layers and stations shown. The solution in the Cali-Tem scenario also got temperature error close to zero °C at the middle and bottom layer at STN. A1, but it has larger temperature error than solution in the Cali-Both at surface layer of STN. A1 and all layers of STN. B1. The solution in the Cali-Vel scenario generally overestimated the water temperature in all locations (i.e., all the surface, middle and bottom layers at both stations). The temperature error of solution in the Cali-Vel scenario is much larger than solution in the Cali-Tem and Cali-Both scenarios in the middle and bottom layer of both stations. The temperature error at most times, for the Cali-Vel scenario, is greater than 0.1 °C."

p.15 L431: *the* calibration...

Response: We will add "the".

Figure caption Fig. 5 L451: in term*s* of...

Response: We will change "term" to "terms".

Figure 6: Values in the darkest hex tiles are almost not readable.

Response: We will change the color of values in the darkest hex tiles to make it readable.

p.19 L536 - 539: This sentence was not clear to me and might require revision.

Response: We will revised the sentence to be: "Outlier or extremely bad solutions are also likely happen for calibration problems where the model output is very sensitive to the calibration parameters (i.e., a small change in model parameters can cause huge changes in the model output that leads to much worse solutions)."

p.20 L553 - 566: Again, the argumentation in this paragraph could be affected by the affect of synthetic observation data as outlined in the first major remark.

Response: We have responded the first major remark in response to R3-1. These hydrodynamic models are built based on physics and knowledge human learned in the past hundreds of years. Hydrodynamic models are not like most of hydrology models that with lots of uncertainty. They of course cannot be try the same as real world situation but they are close representations of the real world situation. We think the investigation on models can provide at least some implications for the investigation people could do with real world data. We will revise L553-566 to limit our language by not making universal statement.

p.20 L573: Remove *these*

Response: We will remove "these".

p.20 L574: suggest *to* have

Response: We will change "suggest" to "suggest to".

p.21 L613: We conclude that the Dynamically Normalized Objective Function *that* we propose

Response: We will revise the sentence to be: “We conclude that the DYNO objective function that we propose”