Response to referee comment Referee #1

We appreciate and would like to thank Referee #1 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

R1-1: This paper presents a new objective function for automatic calibration of 3D hydrodynamic models based on water temperature and velocity data. The case study is a tropical lake, where the authors tested their DYNO+PODS for calibrating a Delft3D model against a subset of previously simulated 3D flow and temperature fields from a run assumed to be the "truth". I strongly appreciate this approach as it provides the full control of the optimization. Moreover, as the authors stressed many times (maybe too many, a lot of repetitions could be avoided), the use of velocity data together with water temperature is not so diffused yet in the field of lakes hydrodynamic modeling but it is crucial to have consistent results and realistic flow fields. Hence I welcomed the author's effort in quantifying how important it is.

Response: Thanks for reviewer's comments.

R1-2: Optimization algorithms as well as suitable objective functions for calibrating complex models are needed in the wide environment of hydrological numerical applications. I enjoyed the reading of the manuscript, which is well written and clearly structured. I appreciated how the authors describe their DYNO and tested its performances. Some clarifications are required, in my opinion, to make the optimization part (which is not the focus of the manuscript but still a key part of it) more accessible to the wide public of HESS, but in general I believe this work is worthy for publication on HESS after some minor revisions.

Response: Thanks for the reviewer's comments. We will revise the optimization part in the revised manuscript (based on this comment and comment at R1-29 and R1-30). Detailed response refers to the response to the comment R1-29 and R1-30.

R1-3: I'd like the authors to consider deepening the analysis on two more aspects which I believe worth a little discussion:

Computational costs: Addressing this aspect is mandatory in a paper on optimization algorithms. The authors make some general considerations here and there, but maybe a dedicated paragraph would be more appropriate. My questions: How many (real) runs of the hydrodynamic model were necessary to get the final solution for e.g. each trial/each configuration of Dyno (temp, vel, both)? What is the computational cost (wall clock time) of these tests? e.g., how much time compared to the error?

Response: Our study set the same evaluation budget (i.e., the maximum number of hydrodynamic model runs) for each trial and calibration scenario (i.e., Cali-Tem, Cali-Vel, and Cali-Both). The maximum number of hydrodynamic model runs in each trial is 192, which is about 8 iterations

with 24 evaluations in each iteration. The result indicates that 8-iterations is a sufficient budget as the calibration progress plot in Figure R1 shows the optimization experiments almost converged in the last few iterations.

The computational time of one simulation takes about 5 hours on a windows desktop with CPU Intel Core i7-4790. However, when running 24 simulations simultaneously on a multi-core platform, the computational time gets longer because of the limited cache memory resources (as discussed in Xia and Shoemaker (2022)). Cache memory is a small amount of much faster memory than main memory. The wall-clock time for one iteration with 24 cores simultaneously running is about 12 hours if using the default process scheduling of the nonuniform memory access (NUMA) multi-core system. We used the mixed affinity scheduling proposed by Xia and Shoemaker (2022), and the wall-clock time is reduced to about 8 hours per iteration. The mixed affinity scheduling changed the default affinity setting by setting a hard affinity on the simulation of each PDE model (i.e., fixing the process of each PDE simulation to one core). This approach proved to be efficient for memory usage and reduced the simulation time. More details about the mixed affinity scheduling and the NUMA system can be found in the study of Xia and Shoemaker (2022). Hence, the wall-clock time of each trial takes about 64 hours (8 iteration×8 hours/iteration).

Xia, W., & Shoemaker, C. A. (2022). Improving the speed of global parallel optimization on PDE models with processor affinity scheduling. Computer-Aided Civil and Infrastructure Engineering, 37(3), 279-299.

We will add a separate paragraph in the revised manuscript as suggested by reviewers to discuss the computational cost of the experiments and put the calibration progress plot (Figure R1) in the supplementary materials.

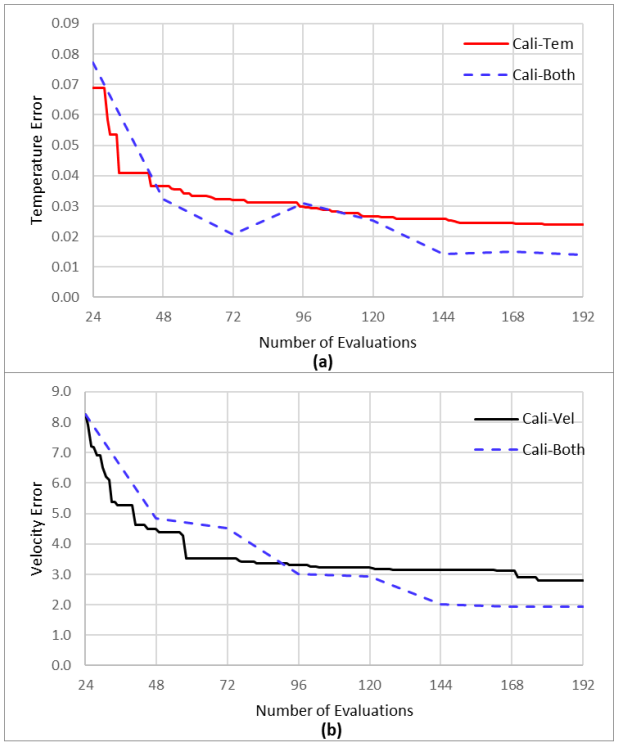

Figure R1. Calibration progress plot of PODS in Cali-Tem, Cali-Vel, and Cali-Both. The best solution found so far in average of three trials is plotted with the number of evaluations.

R1-4: Application to real data from observations: The authors auspicate that future users will test the DYNO against observations and so do I. So my questions are: Are there any constraints in the time/space frequency of the observations? In order to calibrate e.g. their Delft3D lake model to some temperature profiles and some current measurements, how should this data be? E.g. should temp and vel be simultaneous/in the same locations/same depths? As far as I understood, this is indeed the case of the data used in the authors' application, but this is not that common in standard monitoring schemes, where data are sparse and often not simultaneous. So basically, does this optimization (I guess this applies to PODS rather than DYNO) handle sparse observation? Did the authors test their DYNO+PODS by changing e.g. the sampling time or the number of locations of their "truth"? I guess the more data the better, but I'd greatly appreciate some discussion on this. Is there an optimal number of locations/time frequency which gives a satisfactory calibration?

Response: The reviewer is right that we used the temperature and velocity data at the same locations and depths. We are not sure what the result would be if the temperature and velocity data were collected at different places. In reality, we think it is possible that people do measurements of temperature and velocity at the same locations (or locations sufficiently close to each other). We agree that people might measure temperature and velocity at different temporal frequencies (for example, the observation data of one variable might be sparser than that of another). In general, there is no problem in using PODS on problems with sparse observation in the technical aspect. PODS can handle sparse observation. And we don't require the observation of temperature or velocity to be at the same location or with the same frequency since the error of temperature and velocity are calculated separately.

DYNO also does not require the same number of locations / same time frequency for different calibrations constituents. Regardless of number of locations / sampling frequencies, DYNO normalizes the objective function of each constituent dynamically, to allow equal weight to each constituent in the calibration process. In the calibration setup of this study, we believe DYNO has an advantage since it can dynamically adjust the weights between the error of temperature and velocity. In cases when the locations and time-frequencies are vastly different for the two variables, it may be reasonable to introduce custom weights to give more weightage to one constituent. For instance, if velocity observations are limited in both space and time, a custom weight could be introduced in DYNO to reduce overall weightage of velocity in calibration.

We don't think there is an answer to the optimal number of locations and time-frequency. The frequency needs to be small enough to capture the time variation (e.g., diurnal variation or seasonal variation). The number of locations seems dependent on the geography or the reservoir's shape. The location should be enough to capture the spatial variations. The number of locations might also depend on the budget available to do these measures in real practice.

We will add the above discussion into the revised manuscript.

Attached are few minor comments/suggestions to improve the paper.

R1-5: Below few minor comments/suggestions to improve the paper. I strongly recommend the authors to revise the English language as I found some typos (missing spaces, wrong singular/plural verbs) and some passages to be improved, especially in the Abstract, Introduction and Methods. Figures are ok but some "transparency" boxes from the png (my guess) are visible so please consider improving the quality or changing the figure format.

Response: Thanks for the reviewer's comments. We will revise the English language based on reviewer's detailed comments below. We will do a careful proofreading for the revised manuscript. We will also improve the quality of the figures by removing these "transparency" boxes.

R1-6: Title: Please synthetize the title: my suggestion: A Novel Objective Function DYNO for Automatic Multivariable Calibration of 3D Lake Models

Response: Thanks for the reviewer's suggestion. We will change the title to be: "A Novel Objective Function DYNO for Automatic Multivariable Calibration of 3D Lake Models."

R1-7: L19-20: "by comparing the result of using DYNO to results of calibrating to either temperature or velocity observation only" please rephrase

Response: We will revise the sentence to be: "by comparing the calibration result obtained with DYNO to the result obtained through calibrating to only one variable (i.e., temperature or velocity)"

R1-8: L27-33: please revise the English form and make it less general. An example: "Hydrodynamic models simulate the hydrodynamic and thermodynamic processes in lakes and reservoirs": not really, "hydrodynamic models" is a very wide definition for models that can be used to simulate either hydrodynamics only or hydro-thermodynamics (as for Delft3D), and to different water environments, not only lakes. This is just a formal comment and applies to the entire Introduction, please avoid generalized and rough statements as well as repetitions.

Response: We will revise the sentence to be "Lake hydrodynamic models simulate the hydrodynamic or thermodynamic processes in lakes and reservoirs"

R1-9: L30: The authors say that hydrodynamic models simulate specific water quality variables. What do they mean with "specific water quality variables"?

Response: We want to say hydrodynamic models are often built to support water quality modelling of variables such as nutrients, toxins. We will rephrase the sentence to be: "These simulation models (e.g., hydrodynamic modelling) play a critical role in managing water bodies (e.g., rivers, lakes, and coastal areas), as they are built to support the simulation of the spatial and temporal distributions of specific water quality variables (e.g., nutrients, toxins), and to study the response of a water body to different future management scenarios."

R1-10: L46: "some" water variables. I'd say all of them! If the model is 1D, all variables will be 1D. Also in this case, please be more precise: "spatial" is very general, 1D models typically consider the vertical dimension, so what they don't provide is the horizontal spatial distribution and in general they can't capture the 3D processes (e.g. circulations, 2D waves…).

Response: We will revise the sentence to be: "However, one-dimensional models are unable to simulate the horizontal spatial distribution and cannot capture the 3D processes, and thus may not be suitable for certain studies."

R1-11: L66-75: The authors could be interested in reading this work https://doi.org/10.1016/j.envsoft.2021.105017 and references therein where some of the issues mentioned in this paragraph are addressed in a manual calibration of Delft3D in a lake.

Response: Thanks for reviewer's recommendation. The study mentioned by the reviewer is interesting and relevant. The study used different sources of temperature data (from in situ observations, multi-site high-resolution profiles and remote sensing data) to compensate for unavailability/scarcity of velocity measurements. This is a practicable approach when there is no velocity data available and there are such different sources of temperature data available. In cases, there is no high-quality remote sensing data (for example because of cloud) or a large amount of high-resolution profiles of temperature measurement it is still challenging to verify the spatial simulation of hydrodynamic quantities. We will discuss this study in our revised manuscript.

R1-12: L85-90: a little confused, please revise English form.

Response: We will revise the sentence to be: "Hydrodynamics models predict the velocities throughout the water body. These results are important to understand the spatial distribution of water quality problems in sizeable lakes. For the purposes of model calibration it is useful to know whether efforts to measure velocity directly are justifiable if temperature data is already available. We will examine the extent to which direct measurement of velocities justify the extra effort by giving more accurate results for hydrodynamics models. We will also look at the error of the spatial distribution of hydrodynamics associated with calibrating to temperature only, which is rarely studied in literature."

R1-13: L100: etc -> among others; desire -> require

Response: We will replace "etc" with "among others" and replace "desire" with "require".

R1-14: L108: "A key challenge for automatic calibration of multi-variable calibration problems is in defining a suitable objective function to calibrate multiple variables simultaneously": please remove some "calibrations", e.g.: A key challenge for automatic calibration of multi-variable problems is in defining a suitable objective function.

Response: We will revise the sentence to be: "A key challenge for automatic calibration of multi-variable calibration problems is in defining a suitable objective function"

R1-15: L110: varying ◊ vary

Response: We will replace "varying" with "vary".

R1-16: L118: Anticipate MOO to the first time it is mentioned (L113). Does SOO refer to the optimization methods mentioned in lines (105-112)? If yes, please anticipate SOO as well.

Response: We will anticipate MOO to the first time it is mentioned. We will also anticipate SOO at line 109-112: "Traditional approaches typically formulate the goodness-of-fit of multiple variables into a single objective function by adding weights between the goodness-of-fit of multiple variables and solve the problem with single objective optimization (SOO) techniques"

R1-17: Tab.1 check spaces

Response: We will correct the spaces in Table 1.

R1-18: L164: (e.g. calibrating temperature and…) we got that the authors are dealing with multi-variables problems and in particular with temp and vel. Please revise the paper critically and remove repetitions.

Response: Thanks for reviewer's comments. We will revise the paper critically and remove repetitions.

R1-19: Sect.2.3: I see that the point of this work is not the simulation of one specific case study, but since the name of the section is "Study site and data" the authors could at least include the name of the lake and a few morphological characteristics (e.g. where it is located, how deep and large it is) and then refer to Xia et al. 2021 for all other details. Also what year was simulated should be reported for completeness (in the text and in fig. 1).

Response: We agree with the reviewer that it would be better to provide the name of the lake. However, we are constrained by the local agency, which provided us with data, from releasing the lake name, including other confidential information such as reservoir locations and inflow and outflows. We appreciate if the reviewer understands our situation. But we are permitted to mention the depth and the surface of the reservoir. Hence we will add these information in the revised manuscript.

R1-20: L189: just a curiosity, why are the names of the station A1, B1-4? Does this A1 station mean something different than the others?

Response: Station A1 is the station where the real measured temperature data was used for the calibration of the hydrodynamic model. While there is no real observation data at other stations.

R1-21: L211-212: "the water utilities' employees and consultants": some specification is missing here... maybe Singapore?

Response: As responded to the comment above, we are not allowed to disclose the confidential information such as reservoir locations. So we did not mention the name of the local agency.

R1-22: L234: Secchi depth: what about the space-time variability of this parameter? Delft3D allows to consider both variations, as transparency is not a constant and uniform feature of water. Did the authors consider this possibility?

Response: The reviewer is right that Delft3D allows considering space-time variability of these parameters. We consider time-varying parameters such as Secchi depth, Ozmidov length scale, Dalton number, and Stanton number to be constant mainly because it would be very challenging to calibrate them as time-varying variables. Considering these parameters as time or space-varying parameters will substantially increase the number of decision variables in the optimization. In addition, the reservoir we study is relatively small (with maximum depths of 22 m and a surface of 250 hectares), and it is located in a tropical region where there is no significant seasonal variation. Hence we think it is acceptable to consider them as constant. But we think that considering these parameters' space-time variability into optimization calibration would be an interesting future topic. In that case, new methods for reducing the parameter dimensions are needed (e.g., design some low dimensional controlling parameters to represent the high dimensional space-time variability of these parameters). We will add these discussions at the end of the manuscript.

R1-23: L237: please consider moving here the sentence in lines 228-230.

Response: We will move the sentence in lines 228-230 to Line 237 and revise it to be: “The last parameter is the manning coefficient, which affects the roughness of the bottom of the lake and a direct impact on velocity.”

R1-24: L241: Model parameter(s) in table caption

Response: We will replace “Model parameter” with “Model parameters”.

R1-25: Tab.2: Why didn't the authors include the coefficient of free convection, whose value greatly affects the modeled temperature? The default value in Delft3D is 0.14 but it strongly modifies the thermal profile when tuned.

Response: We chose the calibration parameter set as suggested by the local experts on lake modeling (for the region of the study site) and they did not suggest the coefficient of free convection as an optimization parameter. As the reviewer pointed out, this parameter affects the thermal profile, and could be considered as a parameter to be optimized in future.

R1-26: L261: Have the authors considered normalization of RMSE by the standard deviation instead of the mean and why did they eventually chose the mean as normalization factor?

Response: We did not considered normalization of RMSE by standard deviation. RMSE could be normalized with standard deviation or mean. We just chose one of them (in our case we use the mean).

R1-27: L272-273 and Eq.11: It looks to me that there is a power of 2 missing in eq.11. The referred papers (starting from Beletsky et al. 2006) present this formula with the module of the difference between observations and models (at the numerator) and the module of the observations vector (at the denominator) both to the power of 2, and then everything under the square root. If the authors prefer to use the Euclidean norm instead of the module it is fine with me, as ||x||_2 = |x| in R^2, but the power of 2 should be maintained anyway, right? I checked the codes uploaded on github and I see a **2 in the computation of the Fourier norm, but please double check the equation, the code and the references.

Response: The reviewer is right that there is a power of 2 missing in eq.11. We will revise eq.11 in the revised manuscript.

R1-28: L280: “being summed them” remove them

Response: We will remove them.

R1-29: Sect.2.6.1. This paragraph seems like an advertisement of PODS and is highly technical. I'm not sure it is adequate to the wide audience of HESS. The authors could consider limiting the acronyms to those that are really needed (e.g. do we really need to know that “DYCORS inherits the dynamic coordinate search idea from DDS (Tolson and Shoemaker, 2007) to improve its effectiveness and efficiency for high dimensional problems”?) and try to clarify a bit. What does RBF surrogate mean? Maybe the authors could consider merging sect 2.6.1. and 2.6.2. and try to explain things in easier terms (and referring to publications for high-technical details), eventually splitting some very long sentences.

Response: Thanks for reviewer’s comments. We will revise the section 2.6 by merging section 2.6.1 and 2.6.2 as suggested by the reviewer. We will rewrite the text by explaining things in easier

terms and add more details than we previous did. The changes including the changes including more details about RBF surrogates and how it is used in the optimization.

R1-30: Figure2 and commenting text: two aspects are not completely clear to me: 1) how does the RBF surrogate model interact with the new fitness F and how does it then communicate to workeri such that the latter can run a new hydrodynamic simulation? I believe that lines 318-320 are crucial here. The authors could consider expanding these two lines as it seems to me they are taking for granted too many things. 2) how are workers-p related one another? Are all steps M1-M6 performed independently on each processor? If yes, how do they communicate at the end to define the best solution found? If not, do M1-M6 steps consider all trials from workers-p within ψ? And how does M6 discriminate to which worker-i it should communicate the new X?

Response: After one iteration is finished (e.g., the simulation of the *P* hydrodynamic simulations). The objective function value $F(\boldsymbol{X}|\boldsymbol{K})$ of all evaluations in ψ (including the newly finished evaluations in this iteration and previous iterations) are recalculated. The RBF surrogate model is rebuilt with the new objective function value $F(\boldsymbol{X}|\boldsymbol{K})$ that are calculated based on Eq. (3). The newly built RBF surrogate model is then used for selecting the evaluation points for the next iteration. In PODS, the algorithm first generates a large number of candidate points around the best evaluation found so far. Then *P* evaluation points are selected from the candidate points using a Surrogate-Distance Metrics discussed in the PODS paper. The metrics consider the approximated objective function value based on the surrogate model and the distance of the evaluation points from the evaluated points. We will revise lines 318-320 by adding more explanations.

To reviewer's question 2), the tasks of *P* workers are independent of each other (i.e., each worker evaluates one hydrodynamic model). The steps M1-M6 are master's tasks, performed after the tasks of *P* workers are finished in each iteration. So each worker sends back to the master the results of one simulation (i.e., temperature and velocity error of one evaluation). Then the master adds all these newly obtained results to the history list ψ, which contains the results of all evaluated points in previous iterations. M1-M7 only need to be performed on one processor. The jobs of M1-M7 is to generate P evaluation points and then distribute the P evaluation points to P processors for the workers' tasks. The master M7 distributes the P evaluation points randomly to the P processors so each processor will get an evaluation point and send back the evaluated results to master. We will modify the Figure 2 as below to make it clear.

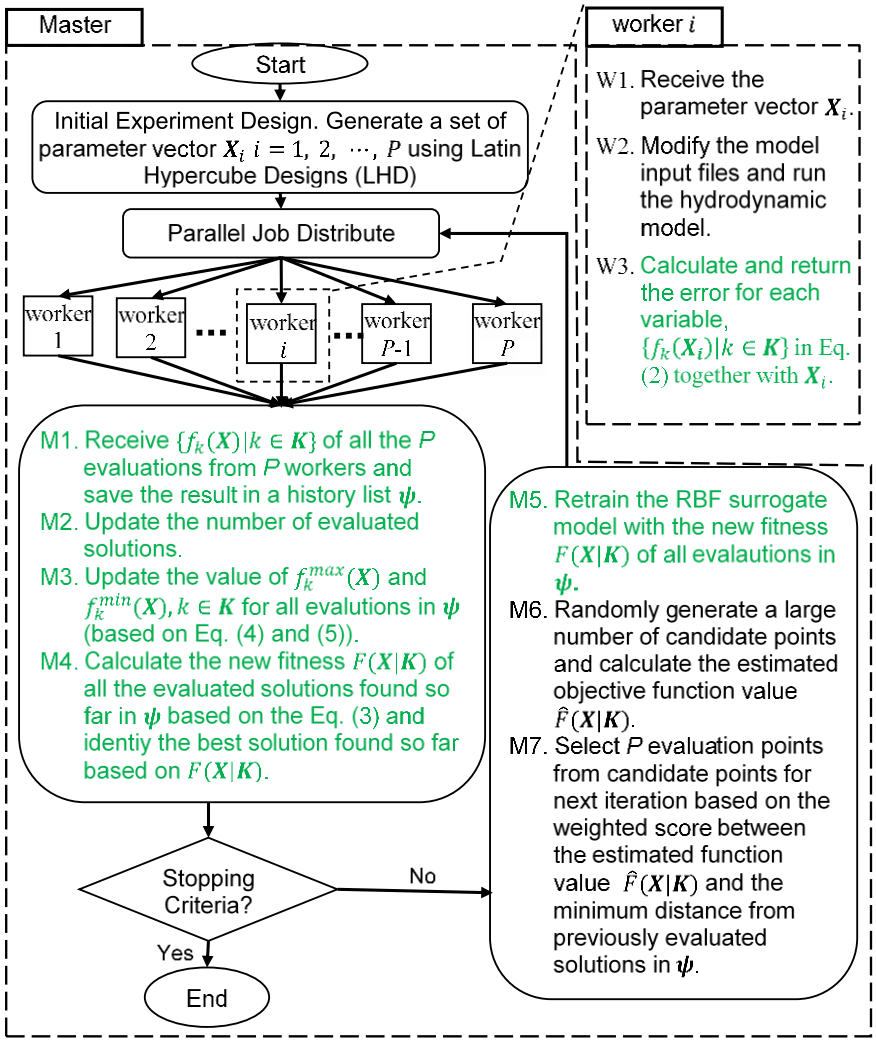

Figure R2 (Revised from Figure 2 in original manuscript). Diagram of the implementation of DYNO with the parallel algorithm PODS. *P* is the number of processors available. The green texts (i.e., steps W3, M1-5) are changes made on PODS to incorporate DYNO. The rest part follows the original PODS method.

R1-31: Table 4 and commenting text: a "good model" range for Fn should be between 0 (perfection) and 1. How do the authors comment such large values of Fn? Their best solution Cali-borh gives almost 2, while the Cali-Vel, which should minimize only Fn, gives almost 3!

Response: The value in Table 4 is the sum of Fn value at multiple stations (in total 12 locations). So the Fn value at one location in average is about 0.167 for the best solution in Cali-Both scenario and 0.25 for best solution in Cali-Vel scenario. Hence it is not a large value of Fn. We will add these explanations in the manuscript.

R1-32: L395: missing dot before Figure3?

Response: We will add the missing dot.

R1-33: L421: overestimated ◊ overestimation

Response: We will replace "overestimated" with "overestimation".

R1-34: L433-434: Latin Hypercube Designs (LHD) is mentioned here for the first time. Please specify what it is used for in the methods section.

Response: We will introduce the Latin Hypercube Designs (LHD) in the method section (i.e., section 2.6.2).

R1-35: L456: "only the best solution in each of the optimization iterations are plotted" ◊ is plotted

Response: We will replace "are" with "is".

R1-36: L462 please change the asterisk with standard math notation

Response: We will change "$\boldsymbol{X}_i, i = 1, \dots, 3 * N_{max}$" to "$\boldsymbol{X}_i, i = 1, \dots, 3 \times N_{max}$"

R1-37: Figure5 and 6 Please make the axis labels consistent between the two figs. fvel(X) and ftem(X) should be fine for fig. 5 (if I got correctly that only the best K is plotted), while ftem(X|K) and fvel(X|K) should be fine in fig. 6

Response: Thanks for reviewer's suggestion. We will revise the Figure 5 as the reviewer suggested.

R1-38: Figure 6: please improve the readability of the number inside the darkest hexagon by either changing the darkest color or modifying the color of the number (e.g.) yellow. Please note that there is a missing C in (c) in the figure legend.

Response: Thanks for reviewer's comment. We will modify the color of the number inside the darkest hexagon and also add the missing "C" in the figure legend.

R1-39: I would have expected the darkest hexagons to be the closest to the origin. Why isn't it like that? What is the hexagon containing the final solution? The authors could consider highlighting it e.g. with a bold colored contour line.

Response: We would like clarify here that Figure 6 represents the joint distribution of the i) Velocity Error and the ii) Temperature error components of DYNO for "all simulations evaluated" during the different calibration scenarios. Hence, Figure 6 is a representation of the optimization search dynamics for the three different objectives analyzed (i.e., Cali-Tem, Cali-Vel and Cali-Both). Figure 6 shows that when calibrating to only temperature or velocity, the search of the optimization only considered the error of one variable and ignored the error of another variable. Hence the darkest hexagon (i.e., the concentration of error distribution) is expected to be close to one of the coordinate axes instead of the origin (e.g., in Figure 6(a) more solutions are found with only better Temperature error, hence darker hexagons are close to vertical axis). When calibrating to both temperature and velocity, the darkest hexagons are close to the origin but not necessarily the closest. This could happen for many reasons. For example, the solution space is multi-modal, and many solutions have the same error around the value of the darkest hexagons. It could also happen that the algorithm searched more around the region that is not close to the best solution. In general, we think it is true that the best solution is the closest to the origin when calibrating to both temperature and velocity, but we don't think it is true that the darkest hexagons must be the closest to the origin.

R1-40: L485: represent (remove s)

Response: We will remove "s".

R1-41: L534: please correct, which one is better than the other? N2 better than N1?

Response: We will correct the sentence to be "…with DYNO-N2 might be better than with DYNO-N1"

R1-42: L537: like◊ likely?

Response: We will revise the sentence to be: "… also likely happen for…"

R1-43: L565: helps to improve the calibrate of ◊ calibration

Response: We will replace "calibrate" with "calibration".