Response to referee comment Referee #2

We appreciate and would like to thank Referee #2 for taking the time and effort to read our manuscript and providing some valuable feedback on the manuscript. 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

Overview:
R2-1: The manuscript is well-written and clear in its intent and results. It is close to publishable if the authors place their approach to calibration in the appropriate context. I am classifying this as "major revisions" because I think the issues are important, but they should not necessarily take a lot of work to implement.

Response: Thanks for reviewer's comment. We have carefully considered the reviewer's comments and will revise the manuscript accordingly based on the comments.

Specific Comments

R2-2: 1. I would have liked to see the introduction have a little that explains the context for the authors choice to test their calibration against a calibrated model rather than against observations. This choice gives them more data to compare against, but with the drawback that the "true" data are biased in exactly the same way as their calibrated results. This is an acceptable, but limited approach -- acceptable because it is useful in understanding and illustrating the new calibration method -- but limited in that it cannot be used to say anything about what might occur when compared to real world data. This idea is emphasized in comment 2 below.

Response: We don't have real observation for the velocity data, and we found in literature it is also rare that people calibrate to both temperature and velocity. This is why we did our analysis with the hypothetical dataset. Using a calibrated model to generate synthetic data is thus the plausible alternative we used to generate observation data. We agree that this approach has limitations. However, since, hydrodynamic models are built based on physics and human-acquired knowledge, they are close representations of real-world situations. Hence, we think that synthetic data generated from calibrated models is a reasonable alternative for real-world data, for this study. We will revise our introduction and text to explain the context why we do investigation, not on real observations, but on synthetic data. We agree that any conclusions drawn from this study should be considered cautiously and hence, will also soften our statements about the value of this analysis in a real-world context (we have emphasized in response to comment R2-3 below).

R2-3: 2. I strongly disagree with the penultimate sentence of the abstract, that the study "suggests that both temperature and velocity measures should be used for hydrodynamic model calibration in real practice." Similar language is found elsewhere in the paper. The model "suggests" nothing and we must remain skeptical about the value of calibration in a real world context without direct illustration of its importance. Unfortunately, the authors' methodology does not support this suggestion or any suggestion about real-world practice. The authors are testing their calibration

against results of a calibrated model -- not the real world. The ability to more precisely capture the calibrated model (by a variant of the same model) cannot be used to imply that real world will be also represented more accurately. This is a fundamental confusion of "precision" -- how close are my answers to grouped together, with "accuracy" -- how close do my answers reflect the real world. The authors have not included any comparisons of their model to real-world data hence they cannot make any statements or suggestions about likely accuracy in representing real-world phenomena.

Response: As responded to comment R2-2, we don't have real observation for the velocity data, and we found in literature it is also rare people calibrate to both temperature and velocity. That's why we did our analysis with the hypothetical dataset. Reviewer holds the view that the model "suggests" nothing. However, the physical model we use is established with scientific knowledge and the law of physics. Hence, it is reasonable to do analysis on synthetic observation data generated from a hydrodynamic model when we do not have real observations available. We agree with the reviewer that we should remain skeptical about the model results. We have revised the sentence to be "Our study implies that in real practice both temperature and velocity measures might need to be considered for hydrodynamic model calibration." We will also revise other similar statements in our manuscripts.

R2-4: 2. If the authors want a stronger paper, they can either compare to real-world observations in a different time period as a classic validation test, or they can examine some important phenomena that are not directly included in the observational data set. For example, the timing of global overturns of the lake are arguably an important phenomena that can be computed from real-world observed data -- the model results for the different calibrations in predicting timing of overturns could be analyzed. I may be wrong, but I suspect that the differences between the various models may not be as significant when compared in their prediction of a large-scale phenomena. The paper would still be publishable, but future researchers would be able to see that control of model biases are likely more important than calibration in capturing real world behaviors.

Response: We don't have data at a different time period to do this validation test. We don't have real observation data for temperature and velocity either. This is why we use the hypothetical dataset for the investigation. The reviewer mentioned the examination of important large-scale phenomena such as overturn. First, we have to mention that this is out of the scope of our study. The model we built for the tropical reservoir is not used for studying overturn. Our data is based on shallow lake in a location with almost no seasonal variation in ambient temperature. Hence, we would not expect overturn to be a significant event in our data set. Lake overturn (e.g., enhanced mixing of upper and lower water levels due to a temperature gradient caused by seasonal air temperature changes) is a complex phenomenon, so one would expect there is an advantage to having both temperature and velocity data information to model this overturn situation. Since the focus of our argument is on obtaining some velocity data rather than entirely relying on temperature data, this argument obviously is valid for lake overturn. There does not seem to be a need to lengthen the paper by adding a section on lake overturn. In addition, we think, to study such large-scale phenomena, it might not be necessary to build a complex 3D model since a 1D or 2D model can do the same job. There are many other purposes of building a 3D model, such as supporting water quality modelling (e.g., the horizontal and spatial distribution of temperature or other water quality parameters like nutrients or toxins). In such a case, the correct temperature and velocity modeling is very important.

R2-5: 3. The authors should include a comparison to an uncalibrated run (using conventional default values).

Response: We will include the comparison to an uncalibrated simulation in the revised manuscript.

R2-6: 4. The authors should discuss the choice of calibration parameters -- how were they chosen, why were they chosen, and what does it mean to hold these as constant parameters. Arguably, a sensitivity study should have been done prior to choosing the calibration parameters. I am somewhat concerned about whether it is physically meaningful to calibrate as fixed parameters values that arguably depend on time-varying physics (e.g., Secchi depth, Ozmidov length scale, Dalton number, Stanton number).

Response: We selected the parameter values by discussing with an expert who is familiar with the Delft 3D simulation model, and he suggested the parameters we optimize. We did not carry out a sensitivity analysis because it requires many simulations. For an expensive simulation model, this means you need to reduce the number of optimization evaluations given a limited total budget for sensitivity analysis and optimization evaluation. Sensitivity analysis (especially global sensitivity) is expensive to do for our problems. We have nine parameters, and each simulation takes about 5 hours to run. Cheaper sensitivity analysis like OAT (one at a time perturbation of parameter value) depends on one's guess of the probable best solution, around which the sensitivity should be done. The sensitivity is the first partial derivative of f(X) with respect to the parameter evaluated at that probable best solution. For example let the function be $f(x_1, x_2) = x_1 + (x_2)^2$. If you assume the best solution is x_1=4 and x_2=2, then the most sensitive parameter is x_2. If you assume the best solution is x_1=1 and x_2=1/4, then the most sensitive parameter is x_1. So this shows OAT is not necessarily reliable for this kind of problem with lots of parameters and a limited computational budget. Discussion of sensitivity analysis methods would add many pages to the manuscript, and we are trying to shorten it. Sensitivity analysis is also not the topic that we are focused on. So here we are using the reasonable practice of getting estimates of sensitive parameters by talking to an expert who has used the model on similar lakes.

In terms of time-varying parameters such as Secchi depth, Ozmidov length scale, Dalton number and Stanton number, we consider them as 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 22 m and the surface 250 hectare) and is located in a tropical region where there is no significant seasonal variation. Hence, we think it is reasonable to considering them as constant. But we do think that considering the space-time variability of these parameters into optimization calibration would be an interesting future topic. In that case, new methods on how to reduce the parameter dimensions are needed (e.g., designing some low dimensional controlling parameters, like curve number in hydrology (Bartlett et al. 2016), to represent the high dimensional space-time variability of these parameters). We will add these discussions in the end of the manuscript.

Bartlett, M. S., Parolari, A. J., McDonnell, J. J., & Porporato, A. (2016). Beyond the SCS-CN method: A theoretical framework for spatially lumped rainfall-runoff response. Water Resources Research, 52(6), 4608-4627.

R2-7: 5. The authors should provide some discussion about the physical meaning of the errors in the results. For example, the Ozmidov length scale in the "true solution" is 0.015, but the "best" calibration has more than double this value; at the same time, the background vertical diffusivity in the calibration is about 3/5 of the true value and the Secchi depth is overestimated by 13%. These all exert controls on vertical mixing, and the wide disparities for different calibration methods makes me question as to whether calibration can really be effective to capture the complex, time-varying behaviors of vertical mixing in a stratified system with the given turbulence model. I do not expect the authors to solve such a problem, but I think they should discuss the implications. Replacing Table 5 with a well-constructed bar graph (with appropriate normalization) would help the readers see which parameters are doing all the work.

Response: It is always challenging to explain the physical effect of parameter error when multiple parameters are being calibrated simultaneously. The relation between parameter value and the simulation results is nonlinear, and it becomes more complex when multiple parameters interact with each other. The reviewer has quantified the difference in "actual" parameter value and "best" parameter value obtained from calibration as a percentage. This might give the impression that the solution is very inaccurate. We are not sure if it is meaningful to look at the parameter error in percentage since some of the parameters value are very small; hence it might lead to a large percentage value when dividing the parameter error by the parameter value. In addition, the solution is obtained with a limited computing budget. Hence it is difficult to find the exactly true solutions in such a limited computing budget. We think it might be more meaningful to visually compared parameter sets values of different calibration scenarios with the true parameter set. Our goal is to show that calibrating to both temperature and velocity gives a better solution than calibrating to one of them. Figure R3 below shows the parameter value of solutions under different scenarios. The parameter value of the solution in calibrating to both is generally closer to the true solution than the other two cases.

The reviewer also questioned whether calibration could really be effective in capturing the complex, time-varying behaviors of vertical mixing. We have analyzed the vertical temperature profiles, the solution in the Cali-Both scenario can almost capture the vertical time-varying temperature profiles of the true solution. In contrast, calibrating to one variable did not fully capture the vertical time-varying temperature profiles. (For example, in April-May for Cali-Tem scenario, in Mar-May, Aug-Sep for the Cali-Both scenario.) We also analyzed the vertical time-varying eddy diffusion and viscosity. The result also indicated that the result in the Cali-Both scenario captures better vertical mixing than the other two scenarios.

In our revised manuscript, we will add the above discussions. We will also replace Table 5 with Figure R3 below. We will add these results about the vertical time-varying temperature, diffusion, viscosity profiles into the supplementary document.

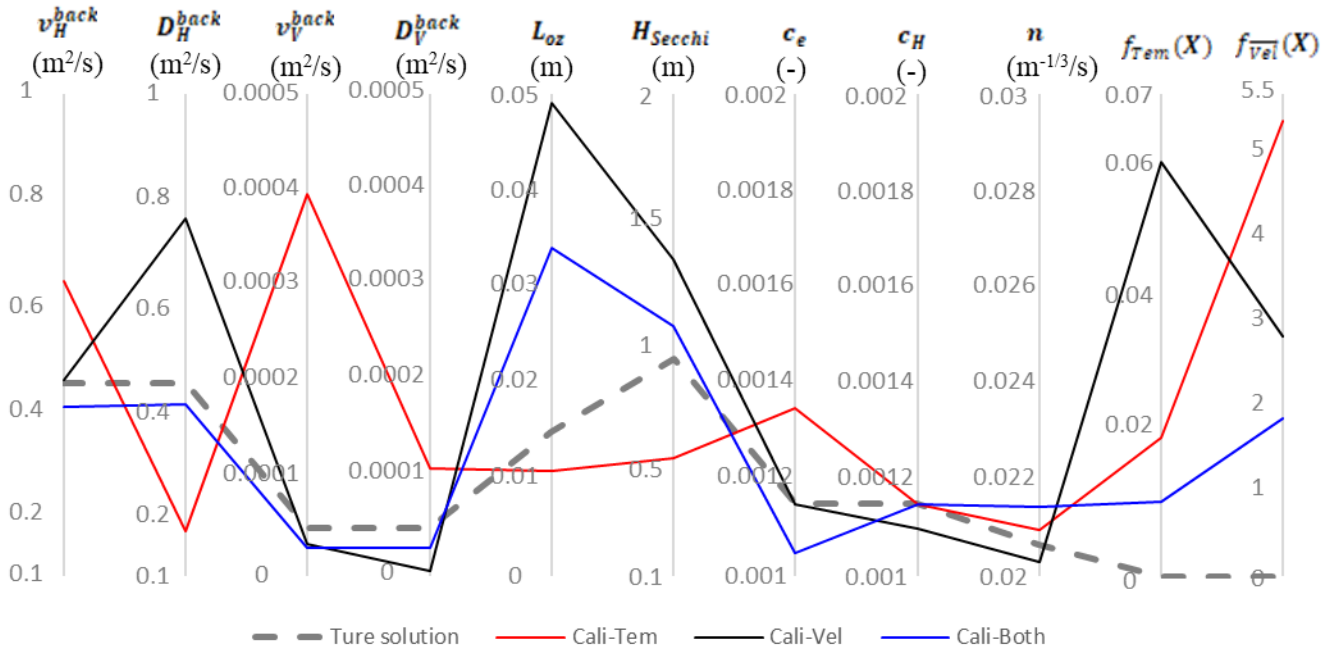

Figure R1. The parallel axis plot for the parameter value and the composite error of temperature and velocity of calibration solutions under different scenarios (Cali-Tem, Cali-Vel, and Cali-Both). True solution defined in Table 2 is given for reference.

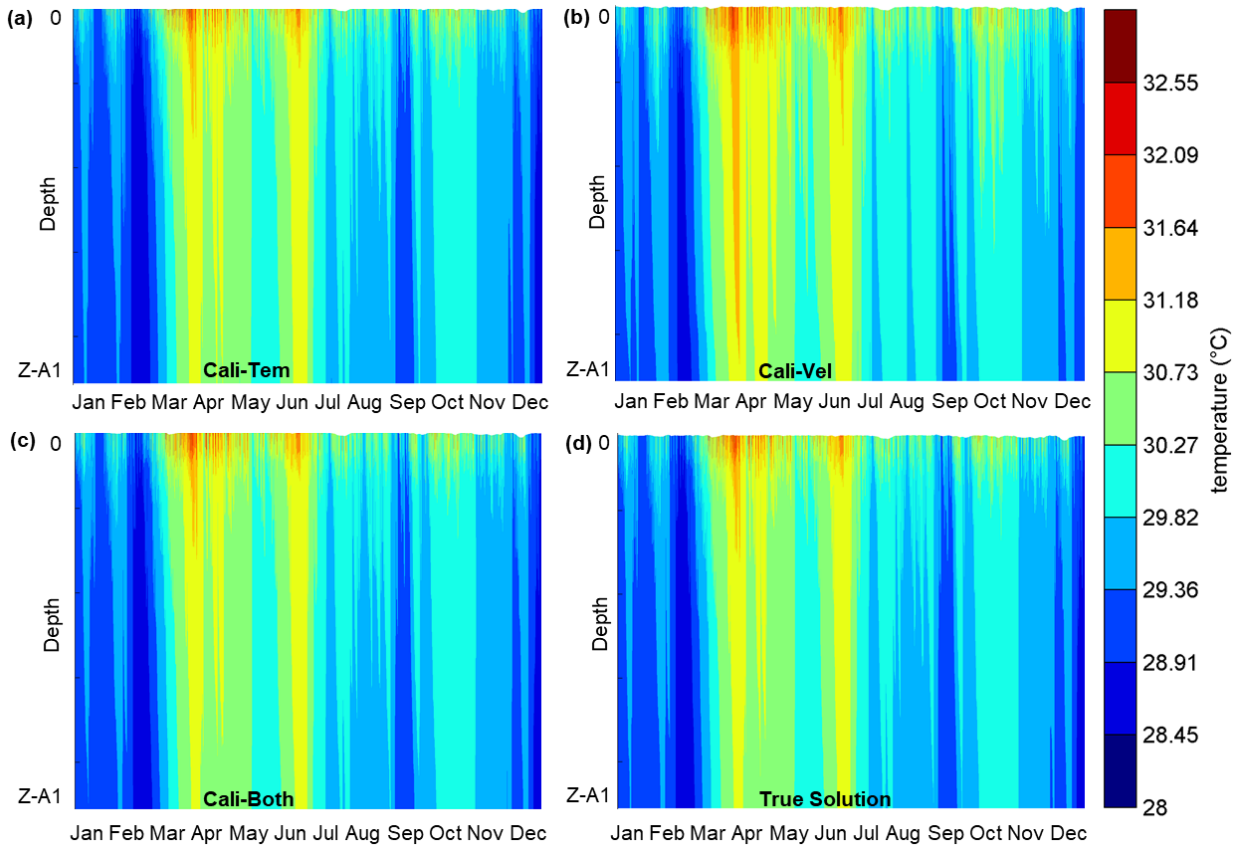

Figure R2 The change of vertical temperature profiles at STN. A1 with the change of time. The result of three calibration scenarios (Cali-Tem, Cali-Vel, and Cali-Both) and the True solution are plotted.

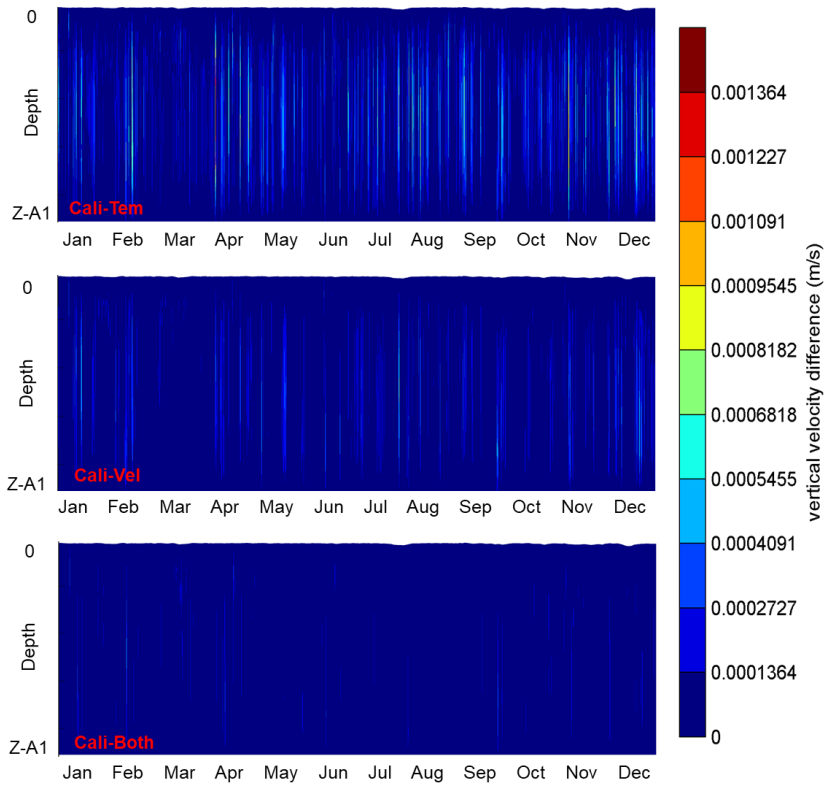

Figure R3 The absolute vertical velocity error at STN. A1 between the calibrated results (in the Cali-Tem, Cali-Vel, and Cali-Both scenarios) and the true solution. The change of absolute vertical velocity error is plotted with the change of time.

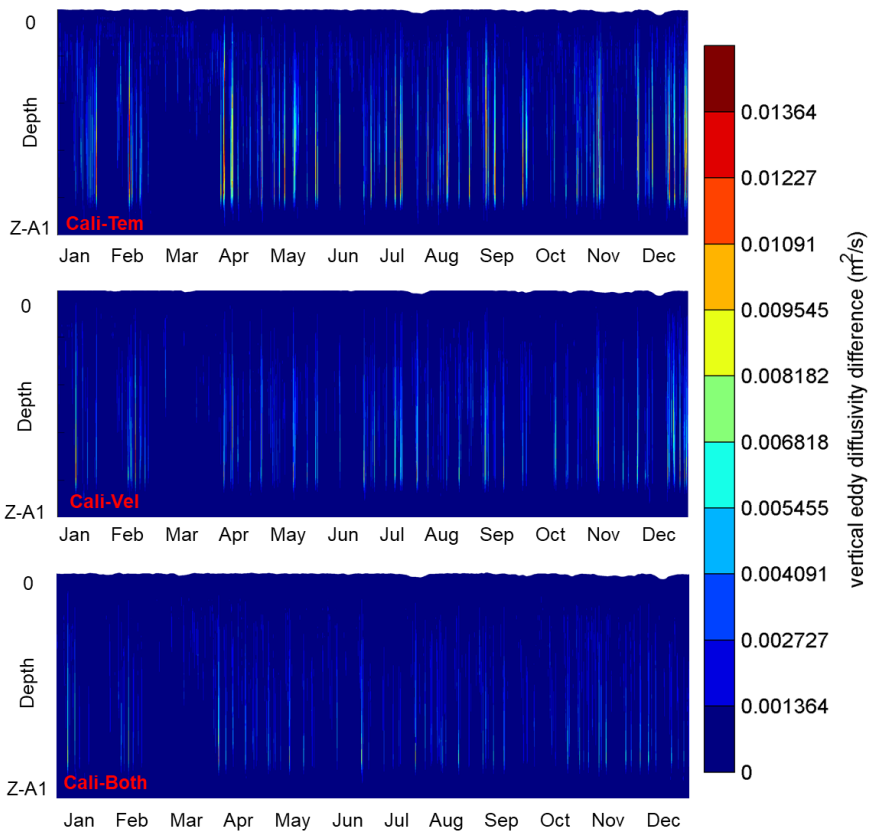

Figure R4 The absolute vertical eddy diffusivity error at STN. A1 between the calibrated results (in the Cali-Tem, Cali-Vel, and Cali-Both scenarios) and the true solution. The change of absolute vertical eddy diffusivity error is plotted with the change of time.

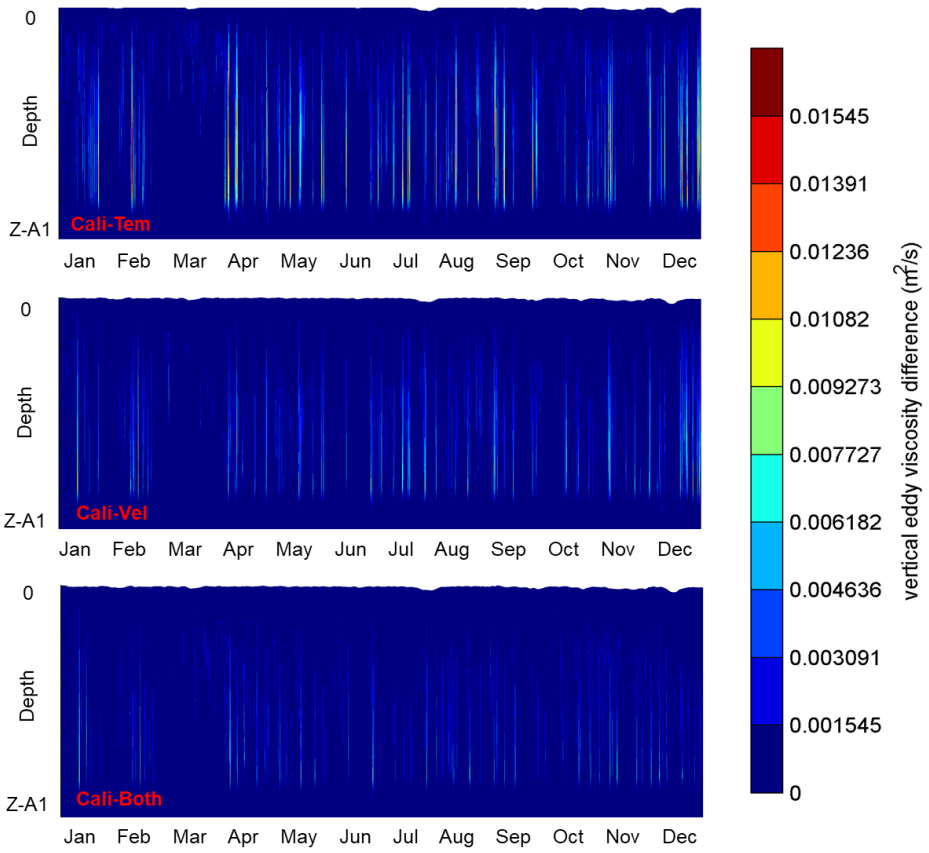

Figure R5 The absolute vertical eddy viscosity error at STN. A1 between the calibrated results (in the Cali-Tem, Cali-Vel, and Cali-Both scenarios) and the true solution. The change of absolute vertical viscosity error is plotted with the change of time.

Path to acceptance:

R2-8: The authors need to carefully limit their language so that the manuscript reflects what can truly be understood from the model and refrain from speculation about real-world behaviors unless they specifically bring new real-world comparisons into the manuscript.

Response: We will revise our language and be skeptical about the model results (as we have responded to R2-1, R2-2).