Reply to Reviewer 1

This summarizes (at some length) efforts to improve the calibration and evaluation of the atmospheric component of the E3SM coupled model. The procedure described seems extremely labor intensive and (frankly) somewhat arbitrary. Nonetheless, the results do show significant improvements and curiously, a reduction in the implied climate sensitivity (assessed via Cess-type perturbations). This is publishable with only minor revisions (as outlined below) and perhaps some condensing to reduce length and repetition.

Reply: We thank the reviewer for the comment. We have considered the reviewer's request to reduce length but respectfully decided to stand by our instinct that to be sufficiently comprehensive about this challenging topic it is important to show all of these diagnostics. We also note that there is precedent in GMD for such technical details for model development and documentation. This paper attempts to carefully document the approach, strategy, and results of model calibration, including the strengths and weaknesses of the approach and results. Furthermore, since most of the recalibration is part of the E3SMv2, this paper serves as a documentation of improvements to E3SMv1 that lead to the new model.

I have two questions that might add to some of the discussion. What is the prospect for automating some of these tests, using ML/AI for instance to reduce the burden and increase the area of phase space tested? I don't have huge confidence that the current procedure will lead to true (local) minima in errors, but I'd like to see this discussed here.

Reply: We have added a paragraph at the end of the paper to discuss the potential use of AI/ML in model calibration:

"It is natural to wonder if an equivalent or superior ESM calibration might have been achievable with less human effort or fewer computational resources via semi-automated machine learning (ML) methods that emulate or expand the workflow outlined in this paper. Indeed, emulating a complex model's parameter sensitivities following human constructed trial simulations to aid model calibration and uncertainty quantification would be an intriguing possibility. Several recent studies have shown successful application of ML methods in model calibration (Cleary et al., 2021;Dunbar et al., 2021;Couvreux et al., 2021;Hourdin et al., 2021). In theory, reinforcement learning (RL) with an appropriately formulated agent-based optimization system could be guided via its loss function formulation with skill metrics that optimize for the same patterns and mean state climate metrics that we prioritized in this study. In practice, however, this ML task faces a fundamental challenge that the cost of an individual agent-reward sample is performing multi-year climate simulations. The workflow outlined in this paper has the considerable advantage that experienced human experts make educated parameter interventions based on assessment of the simulation that discriminates desired effects in a nuanced way and tolerates certain unintended consequences. It is not clear how available ML methods could be infused with analogous physical foresight to make similar decisions, and thus logical to expect they would require more evaluation samples to succeed via brute force. Therefore, experimenting with clever strategies to increase reward density and to integrate physical knowledge from experts in the ML workflow would be a highly worthy long-term challenge."

Secondly, there is a preprint related to ECS in CESM2 (a related model), that has pointed out some odd (possibly erroneous) coding related to the ice-nucleation in that model (CAM6). Does this have any relevance here? https://www.essoar.org/pdfjs/10.1002/essoar.10507790.1

Reply: In E3SM, we do use the cloud ice number limiter (Nimax), but it is set to the in-cloud ice nucleation number. In Zhu et al (2021), they indicated that just removing nimax is not sufficient to lower ECS (and on its own it actually leads to unrealistic features). The key seems to be removing nimax as well as increasing the number of microphysics sub timesteps (which would add computational cost). Given these constraints, we decided not to pursue this but we agree that this could be part of a future sensitivity study.

Minor points:
line 40. "...precise knowledge of ... ERF is not enough". This is a strawman argument. Who has ever said that it was?

Reply: In this study, we find that even though the recalibration does not change the global mean ERFs, the sensitivity of clouds, precipitation, and surface temperature to aerosol perturbations is significantly reduced. This suggests that global mean ERFs are not enough to constrain historical or future climate change. We think this statement is consistent with our findings. We also added a paragraph describing the empirical relationship between ERF and ECS in CMIP models and why it is scientifically interesting to point out that the global mean ERFs are insufficient: "Furthermore, an empirical relation has been shown to exist between the global mean ERFant and ECS in climate models from both the CMIP3 and CMIP5 collections (Kiehl, 2007;Forster et al., 2013). The relationship between ERFant and ECS exists because both values in models are sensitive to simulated clouds. Our tuning strategy specifically targets improving the representation of clouds, and it is worth asking whether these improvements uphold or alter the ERFant -ECS relation. The small difference in ERFant between the EAMv1 and EAMv1P configurations suggests the possibility of a similar small difference in ECS between these two configurations, and yet we find this is not the case."

line 84. The comparison to other ESMs is irrelevant. It the comparison to the constrained range from observations that matters (Sherwood et al, 2020).

Reply: We have replaced "compared to other ESMs" with "compared to estimates based on multiple lines of evidence including process understanding, historical climate record, and paleoclimate record (Sherwood et al., 2020)".

line 114-115. Is there any evidence that the skill scores dervied from a 5 day simulation are correlated to skill scores from a year or 10 year run? Presumably they are not being tested against the same observations?

Reply: We agree with the reviewer that calibration based on short simulations might reach a different configuration than that based on long simulations, because short simulations are intended for understanding the fast physics (e.g., clouds, convections, turbulence) and their local effects, while multi-year simulations include the large-scale feedback. In line 121-124, we stated that "Another limitation is that the short simulations focus on fast physical processes and rapid adjustments. By design, important factors such as slow internal variability of the atmosphere and circulation feedbacks are not considered, so any conclusion drawn from the short simulation ensemble might not be applicable to the calibration of the ESM for climate simulations."

line 120: "in hindsight"? is this referring to the 5-day simulations with EAMv1, or the previous one-at-a-time approach.

Reply: This refers to the 5-day simulations. In Line 120, we stated that "In hindsight, the parameter set selected for the short simulation ensemble during the EAMv1 development was insufficient because parameters not included in the original ensemble were later found to be important."

line 125. There is a big gap between 5 days and 10 years. Is there any assessment of how useful different lengths of simulation might be? For instance 1 year might be a good compromise?

Reply: The length of the simulations depends on the scales that model developers intend to consider in the model calibration. Short simulations with simulation length of a few days are useful for understanding fast physics and adjustments (Ma et al., 2021;Ma et al., 2014;Xie et al., 2012). In this study, we intended to

account for slow internal variability (e.g., inter-annual variability), so multi-year simulations are needed. In the development stage of EAMv1, both 5-yr and 10-yr simulations were performed.

line 128. use the actual times (10 years and 5 days) rather than 'short' or 'long' - relative measures are not very specific.

Reply: Both Wan et al. (2014) and Qian et al. (2018) use "short ensemble" to describe their methods because short simulations are not necessarily 5 days long. Therefore, we believe "short ensemble" is the appropriate term here. For clarification, we revised the text: "The one-at-a-time calibration approach using multi-year simulations and the short simulation ensemble approach using multi-day simulations…"

line 145. "perfect" is too much to ask. But the point about non-uniqueness is important.

Reply: We agree that we do not intend to achieve a perfect model configuration in this study. We stated that "out calibration does not lead to a unique and perfect configuration".

line 160. Why? The authors just spent two pages saying why this was not a good approach!

Reply: We compared the two approaches and discussed their strengths and weaknesses. The purpose of this study is to improve the model fidelity through physics guided tuning and to understand the impacts of parameter changes on the simulated clouds, precipitation, and climate. Therefore, the one-at-a-time approach is appropriate for this study.

line 224. "might"? --> "will"

Reply: We changed the text to "It is logical to expect that increasing model spatial resolution will reduce the impacts of these subgrid effects. Thus, a retuning of these subgrid effects would likely be needed when the model is run at a different horizontal resolution."

line 245. Has the length of these simulations been mentioned?

Reply: The design of the simulations, including the length, is described in Section 2.4.

line 385. How long are these simulations? (line 400 suggests 11 years, but is that just for simulation #5?) In any case, move this up in the text.

Reply: Done.

line 420/Table 6. Add observed values (where available) for comparison (i.e. from CERES, or CALIPSO).

Reply: We have added the satellite observations in Table 6 and added the text "Satellite observations summarized in Stubenrauch et al. (2013) and Neubauer et al. (2019) are also provided but we note that it is dangerous, and can be misleading, to compare model state variables with satellite retrievals without using a simulator since large retrieval and sampling uncertainties exist."

line 438-449. please compare with Cesana et al (2021, doi:10.1029/2021GL094876). The implementation of a CALIPSO simulator should indeed be a high priority. Without a realistic target for LCF this tuning will inevitably be haphazard, but I think it likely that the EAMv1P is more realistic.

Reply: We agree that the implementation of the CALIPSO simulation with cloud phase diagnostics will be very important. Unfortunately, EAMv1 does not have such capability. We have already mentioned this

issue in the manuscript: “While the CMIP5 models tend to freeze liquid condensates at higher temperatures (Cesana et al., 2015;Tan et al., 2016;McCoy et al., 2016), EAMv1 appears to have overcorrected this bias and produced excessive supercooled liquid at low temperatures. Consistent with Zhang et al. (2019), EAMv1_MP increases the T5050. Combining with changes introduced in EAMv1_ZM, EAMv1P produces a much more reasonable T5050 of 254K, which is at the lower bound of the observational estimates. We note that even though Hu et al. (2010) provided an observationally derived LCF-T relationship based on the Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation (CALIPSO) measurements (Winker et al., 2007), EAMv1 does not have the CALIPSO cloud phase simulator (Cesana and Chepfer, 2013) so that a fair comparison is not possible. Evaluating the model LCF-T relationship against satellite observations in a consistent way will be very useful and requires further investigation.”

Cesana et al. (2021) pointed out the need for accounting for snow in the radiation calculation. Since E3SM already account for snow radiative effects, we do not think this is relevant to our study.

figure 4. The authors should add the EAMv1P-CERES map as well, so that the improvemed version can be compared directly with panel a. (also in Figures 5, 6, 9, 10, 11, and 12)

Reply: A primary goal of this study is to assess the impacts of parameter changes on the simulated climate. Comparing different model configurations and the observations does not achieve that goal. Therefore, we believe that showing the differences between the default and the recalibrated model is the best way to present the results.

line 497. It is of course challenging, but I don't think the biggest challenge is the lack of observational data.

Reply: We have removed the sentence.

line 635. This will always be true - it is not a binary situation.

Reply: We changed the text to “While the possibility of compensating biases always exists, our confidence in the underlying physics in the model will be increased if many other aspects are also improved. Otherwise, we are forced to suspect that the model achieves its behavior primarily through compensating biases.”

line 802. This is an odd argument. Who has ever claimed that ERF is sufficient to determine responses? It is precisley the opposite - the major uncertainty (since the Charney report!) has always been in the sensitivity.

Reply: In this paragraph, we summarize our findings by stating that the default model EAMv1 and the recalibrated model EAMv1P produce the same global mean ERFs but their responses of hydrological and surface temperature to aerosols are different. Hence, we find that global mean ERFs are insufficient to understand the response of hydrological cycle and surface temperature to aerosols. The shortwave and longwave contribution to the total aerosol ERF as well as the spatial distribution of aerosol ERF need to be considered. We think this statement is consistent with our findings. We also added a paragraph describing the empirical relationship between ERF and ECS in CMIP models and why it is scientifically interesting to point out that the global mean ERFs are insufficient: “Furthermore, an empirical relation has been shown to exist between the global mean ERFant and ECS in climate models from both the CMIP3 and CMIP5 collections (Kiehl, 2007;Forster et al., 2013). The relationship between ERFant and ECS exists because both values in models are sensitive to simulated clouds. Our tuning strategy specifically targets improving the representation of clouds, and it is worth asking whether these improvements uphold or alter the ERFant -ECS relation. The small difference in ERFant between the EAMv1 and EAMv1P configurations suggests the possibility of a similar small difference in ECS between these two configurations, and yet we find this is not the case.”

line 818+. The comparison to the other models is fine, but the comparison should be with observationally constrained estimates - ie. Sherwood et al (2020), IPCC AR6 Chp. 7 etc.

Reply: We have added Sherwood et al (2020). We did not cite IPCC AR6 because the webpage stated that the documents carry the note from the Final Government Distribution "Do Not Cite, Quote or Distribute".

line 1019. This has never been claimed.

Reply: In this paragraph, we stated that even though the default model EAMv1 and the recalibrated model EAMv1P have the same global mean ERF, they produce different surface temperature response to aerosols and different cloud feedback. Hence, global mean ERF is not a good indicator for historical or future climate change. We think this statement is correct and consistent with our findings. We also added a paragraph describing the empirical relationship between ERF and ECS in CMIP models and why it is scientifically interesting to point out that the global mean ERFs are insufficient: "Furthermore, an empirical relation has been shown to exist between the global mean ERFant and ECS in climate models from both the CMIP3 and CMIP5 collections (Kiehl, 2007;Forster et al., 2013). The relationship between ERFant and ECS exists because both values in models are sensitive to simulated clouds. Our tuning strategy specifically targets improving the representation of clouds, and it is worth asking whether these improvements uphold or alter the ERFant -ECS relation. The small difference in ERFant between the EAMv1 and EAMv1P configurations suggests the possibility of a similar small difference in ECS between these two configurations, and yet we find this is not the case."

Reference

Cesana, G., and Chepfer, H.: Evaluation of the cloud thermodynamic phase in a climate model using CALIPSO-GOCCP, J Geophys Res-Atmos, 118, 7922-7937, 10.1002/jgrd.50376, 2013.

Cesana, G., Waliser, D. E., Jiang, X., and Li, J. L. F.: Multimodel evaluation of cloud phase transition using satellite and reanalysis data, J Geophys Res-Atmos, 120, 7871-7892, 10.1002/2014jd022932, 2015.

Cesana, G. V., Ackerman, A. S., Fridlind, A. M., Silber, I., and Kelley, M.: Snow Reconciles Observed and Simulated Phase Partitioning and Increases Cloud Feedback, Geophys Res Lett, 48, e2021GL094876, https://doi.org/10.1029/2021GL094876, 2021.

Cleary, E., Garbuno-Inigo, A., Lan, S. W., Schneider, T., and Stuart, A. M.: Calibrate, emulate, sample, J Comput Phys, 424, ARTN 109716, 10.1016/j.jcp.2020.109716, 2021.

Couvreux, F., Hourdin, F., Williamson, D., Roehrig, R., Volodina, V., Villefranque, N., Rio, C., Audouin, O., Salter, J., Bazile, E., Brient, F., Favot, F., Honnert, R., Lefebvre, M. P., Madeleine, J. B., Rodier, Q., and Xu, W. Z.: Process-Based Climate Model Development Harnessing Machine Learning: I. A Calibration Tool for Parameterization Improvement, J Adv Model Earth Sy, 13, ARTN e2020MS002217, 10.1029/2020MS002217, 2021.

Dunbar, O. R. A., Garbuno-Inigo, A., Schneider, T., and Stuart, A. M.: Calibration and Uncertainty Quantification of Convective Parameters in an Idealized GCM, J Adv Model Earth Sy, 13, ARTN e2020MS002454, 10.1029/2020MS002454, 2021.

Forster, P. M., Andrews, T., Good, P., Gregory, J. M., Jackson, L. S., and Zelinka, M.: Evaluating adjusted forcing and model spread for historical and future scenarios in the CMIP5 generation of climate models, J Geophys Res-Atmos, 118, 1139-1150, 10.1002/jgrd.50174, 2013.

Hourdin, F., Williamson, D., Rio, C., Couvreux, F., Roehrig, R., Villefranque, N., Musat, I., Fairhead, L., Diallo, F. B., and Volodina, V.: Process-Based Climate Model Development Harnessing Machine Learning: II. Model Calibration From Single Column to Global, J Adv Model Earth Sy, 13, ARTN e2020MS002225, 10.1029/2020MS002225, 2021.

Hu, Y. X., Rodier, S., Xu, K. M., Sun, W. B., Huang, J. P., Lin, B., Zhai, P. W., and Josset, D.: Occurrence, liquid water content, and fraction of supercooled water clouds from combined CALIOP/IIR/MODIS measurements, J Geophys Res-Atmos, 115, Artn D00h34, 10.1029/2009jd012384, 2010.

Kiehl, J. T.: Twentieth century climate model response and climate sensitivity, Geophys Res Lett, 34, Artn L22710, 10.1029/2007gl031383, 2007.

Ma, H. Y., Xie, S., Klein, S. A., Williams, K. D., Boyle, J. S., Bony, S., Douville, H., Fermepin, S., Medeiros, B., Tyteca, S., Watanabe, M., and Williamson, D.: On the Correspondence between Mean Forecast Errors and Climate Errors in CMIP5 Models, J Climate, 27, 1781-1798, 10.1175/Jcli-D-13-00474.1, 2014.

Ma, H. Y., Zhou, C., Zhang, Y. Y., Klein, S. A., Zelinka, M. D., Zheng, X., Xie, S. C., Chen, W. T., and Wu, C. M.: A multi-year short-range hindcast experiment with CESM1 for evaluating climate model moist processes from diurnal to interannual timescales, Geosci Model Dev, 14, 73-90, 10.5194/gmd-14-73-2021, 2021.

McCoy, D. T., Tan, I., Hartmann, D. L., Zelinka, M. D., and Storelvmo, T.: On the relationships among cloud cover, mixed-phase partitioning, and planetary albedo in GCMs, J Adv Model Earth Sy, 8, 650-668, 10.1002/2015ms000589, 2016.

Neubauer, D., Ferrachat, S., Siegenthaler-Le Drian, C., Stier, P., Partridge, D. G., Tegen, I., Bey, I., Stanelle, T., Kokkola, H., and Lohmann, U.: The global aerosol-climate model ECHAM6.3-HAM2.3-Part 2: Cloud evaluation, aerosol radiative forcing, and climate sensitivity, Geosci Model Dev, 12, 3609-3639, 10.5194/gmd-12-3609-2019, 2019.

Qian, Y., Wan, H., Yang, B., Golaz, J. C., Harrop, B., Hou, Z. S., Larson, V. E., Leung, L. R., Lin, G. X., Lin, W. Y., Ma, P. L., Ma, H. Y., Rasch, P., Singh, B., Wang, H. L., Xie, S. C., and Zhang, K.: Parametric Sensitivity and Uncertainty Quantification in the Version 1 of E3SM Atmosphere

Model Based on Short Perturbed Parameter Ensemble Simulations, J Geophys Res-Atmos, 123, 13046-13073, 10.1029/2018jd028927, 2018.

Sherwood, S. C., Webb, M. J., Annan, J. D., Armour, K. C., Forster, P. M., Hargreaves, J. C., Hegerl, G., Klein, S. A., Marvel, K. D., Rohling, E. J., Watanabe, M., Andrews, T., Braconnot, P., Bretherton, C. S., Foster, G. L., Hausfather, Z., Heydt, A. S., Knutti, R., Mauritsen, T., Norris, J. R., Proistosescu, C., Rugenstein, M., Schmidt, G. A., Tokarska, K. B., and Zelinka, M. D.: An Assessment of Earth's Climate Sensitivity Using Multiple Lines of Evidence, Rev Geophys, 58, ARTN e2019RG000678, 10.1029/2019RG000678, 2020.

Stubenrauch, C. J., Rossow, W. B., Kinne, S., Ackerman, S., Cesana, G., Chepfer, H., Di Girolamo, L., Getzewich, B., Guignard, A., Heidinger, A., Maddux, B. C., Menzel, W. P., Minnis, P., Pearl, C., Platnick, S., Poulsen, C., Riedi, J., Sun-Mack, S., Walther, A., Winker, D., Zeng, S., and Zhao, G.: Assessment of Global Cloud Datasets from Satellites: Project and Database Initiated by the GEWEX Radiation Panel, B Am Meteorol Soc, 94, 1031-1049, 10.1175/Bams-D-12-00117.1, 2013.

Tan, I., Storelvmo, T., and Zelinka, M. D.: Observational constraints on mixed-phase clouds imply higher climate sensitivity, Science, 352, 224-227, 10.1126/science.aad5300, 2016.

Wan, H., Rasch, P. J., Zhang, K., Qian, Y., Yan, H., and Zhao, C.: Short ensembles: an efficient method for discerning climate-relevant sensitivities in atmospheric general circulation models, Geosci Model Dev, 7, 1961-1977, 10.5194/gmd-7-1961-2014, 2014.

Winker, D. M., Hunt, W. H., and McGill, M. J.: Initial performance assessment of CALIOP, Geophys Res Lett, 34, Artn L19803, 10.1029/2007gl030135, 2007.

Xie, S. C., Ma, H. Y., Boyle, J. S., Klein, S. A., and Zhang, Y. Y.: On the Correspondence between Short- and Long-Time-Scale Systematic Errors in CAM4/CAM5 for the Year of Tropical Convection, J Climate, 25, 7937-7955, 10.1175/Jcli-D-12-00134.1, 2012.

Zhang, Y., Xie, S., Lin, W., Klein, S. A., Zelinka, M., Ma, P.-L., Rasch, P. J., Qian, Y., Tang, Q., and Ma, H.-Y.: Evaluation of Clouds in Version 1 of the E3SM Atmosphere Model With Satellite Simulators, J Adv Model Earth Sy, https://doi.org/10.1029/2018MS001562, 2019.