We would like to thank the reviewer for positive and helpful comments and suggestions. Our responses and actions are indicated in blue below:

This is an easy-to-read and useful introduction to NCAR's perturbed parameter ensemble (PPE) methodology. The results are interesting and important. However, I think that some additional discussion by the authors could clarify the interpretation of the results.

Major comments:

Lines 77–78: "we initially created 250 different sets of parameter values in addition to the default CESM2-CAM6 setup (total of 251 sets)." How was this number chosen? What is the consequence if only half of the 250 parameters sets are emulated? The paper contains a nice comparison of emulator techniques, but I wonder if an excellent emulator could be thwarted by a sparse sample of parameter sets.

The number of ensembles were chosen based upon a parameter to ensemble numbers factor of approximately 5. This follows a study by Regayre et al. (2023), which used a ratio of about 6.

Regayre, L. A., Deaconu, L., Grosvenor, D. P., Sexton, D. M. H., Symonds, C., Langton, T., Watson-Paris, D., Mulcahy, J. P., Pringle, K. J., Richardson, M., Johnson, J. S., Rostron, J. W., Gordon, H., Lister, G., Stier, P., and Carslaw, K. S.: Identifying climate model structural inconsistencies allows for tight constraint of aerosol radiative forcing, Atmos. Chem. Phys., 23, 8749–8768, https://doi.org/10.5194/acp-23-8749-2023, 2023.

We added this text:
The ratio of number of ensembles to numbers of parameters is ~5.5, close to the ratio of 6 used in Regayre et al. 2023.

The comment regarding the consequence of numbers of ensemble members is an important one. Ongoing work indicates that the number might matter in that it is possible that parts of the parameter space are missed with sparse sampling. It seems that some types of emulator techniques perform better than others when sampling is sparse. Indeed there are many questions related to the construction of a PPE whose goal is parameter estimation. Number of samples is one, another is sampling strategy (LHS vs. CliMA's Ensemble Kalman methods vs. PPE-plus-first-guess-CPE iterative approach). There is also the question of whether to emulate "raw" outputs or some distillation of them. Finally the effect of a probabilistic emulator—whether GP or an ensemble of for example NNs—is also of critical importance. Answering the question of effect of ensemble size should ideally be tackled in the context of addressing some, most, or all of these questions, but that is outside the scope of the current work. However, work is being done on comparing the CAM6 PPE with another PPE with more ensemble members and comparing how they do perform with different emulators.

Line 241: “First, we will illustrate the basic spread across the PPE for several key features of the simulated climate system.” Is this the ensemble spread over all values based on a uniform sample of parameter values within the expert-chosen ranges?

Yes. We have added the sentence:
The ensembles are spread over all parameter values based upon the uniform sampling of the parameter values within the expert chosen ranges. The magnitude of the spread in output values is dependent both on the sensitivity of the parameter, the range of the parameter, and the combinations together with other parameters.

Could you provide some comments on how these spreads should be interpreted? The magnitude of the spread would be expected to depend sensitively on the chosen parameter range. If that range is chosen subjectively, then the spread will inherit that subjectivity. In principle, the range might be problematic because, e.g., we know that some ensemble members produce an unrealistic climate, because the authors state on lines 409–410 that “First, we will illustrate the basic spread across the PPE for several key features of the simulated climate system”. Given this, are the spreads realistic?
E.g., the max ice fall speed factor is 25 times the min value, suggesting that we don’t know the ice fall speed within an order of magnitude. Is this true? To cite another example, the accretion enhancing factor varies by a factor of 100. Is this degree of uncertainty realistic?

The magnitude of the spread in output values is dependent both on the sensitivity of the parameter, the range of the parameter, and the combinations together with other parameters. For example, the ice fall speed parameter, which is sensitive in LWCF (Figure 7), can have the same value of LWCF at the extreme end of the parameter values, depending on the values of the other parameters. But it is correct that certain ensemble members can create unrealistic climate values due to the combination of the parameter values, and some studies remove these members in their analysis (for example Duffy et al. 2023).

However, regardless of whether the ranges of a given parameter are realistic, specification of some a priori range is usually done from the perspective of univariate parameter variation. Because it is typically unknown how parameter perturbations will interact, when performing simultaneous perturbations, it is advantageous to consider wider ranges of parameters, because these will more fully elucidate the model's ability to produce compensating errors. Compensating errors, in turn, help to indicate where independent information on individual parameters (observations, a priori theoretical or laboratory information) may be needed to independently constrain parameters and break the compensation of errors.

Finally, the ice fall speed range is perhaps large. However, knowing that there is a large uncertainty in the bulk ice particle habit the model tries to represent, the uncertainty inherent for this parameter is large.

We added this text in paragraph 3.3:

Some of the chosen ranges are large. However, regardless of whether the ranges of a given parameter are realistic, specification of some a priori range is usually done from the perspective of univariate parameter variation. Because it is typically unknown how parameter perturbations will interact, when performing simultaneous perturbations, it is advantageous to consider wider ranges of parameters. These will more fully elucidate the model's ability to produce compensating errors. Compensating errors, in turn, help to indicate where independent information on individual parameters (observations, a priori theoretical or laboratory information) may be needed to independently constrain parameters and break the compensation of errors.

Minor comments:
Lines 44-46: "in the current study we perturb 45 different parameters, which would require a minimum of 3.5·10^13 (2^45) simulations using OAT if each parameter was tested with only two values in all combinations." How are you defining "OAT" here? 2^45 sample points would fill the entire 45-dimensional volume of parameter space, which would involve perturbing all parameters simultaneously, not one at a time. Perturbing each parameter individually would yield only 2*45 samples, no?
One at a time means that for each new ensemble member, we change only one parameter. Let's say we have 3 parameters. First ensemble uses parameters a,b,c. Next ensemble uses a+1,b,c. Then the next ensemble uses a+1,b+1,c, then a+1,b+1,c+1, then a,b+1,c….. Each new ensemble member only changes one parameter. To fill all parameters possibilities, we here need 8 members (2^3).

Table 1: The max value of DCS is listed as 1.0e-6. Should it be 1000e-6?
Correct, this is fixed.

Line 247: "global, annual mean net top of atmosphere flux balance (TOA: Figure 3H)." The definition of TOA is unclear to me. Does a net downward flux have positive TOA? Or negative TOA? Is TOA different than RESTOM (line 297)?

The TOA is the net flux at the top of the atmosphere, and yes, net downward flux has a positive TOA. RESTOM, on the other hand, is the net flux at the top of the model level. In the model, the TOA fluxes adds an additional layer above the top of the model to provide a more appropriate comparison with satellite observations. We added the sentence:
Note, RESTOM is similar to TOA energy balance, but at the top of the model as opposed to top of the atmosphere.

Line 414: Replace "relative" with "relatively".
Done.

Line 417: Replace "Colombia" with "Venezuela".
We have replaced Colombia with Columbia.