Answers to Referee Ian Enting comments: ***Review of Calibration of a multi-physics ensemble for greenhouse gas atmospheric transport model uncertainty estimations***

We thank the referee for the helpful comments that will improve the manuscript. In the text below, we have tried our best to respond to all the general and specific comments provided by the reviewer.

Comments to Author:

This is a significant study and is appropriate for publication in ACP. However there a few places where the terminology could be clarified.

REF-C1: Overall, I think the term "selection" or "down-selection" is preferable to "calibration " for the process that is being used.

Author-C1: We agree with the reviewer that our method is basically a selection of ensemble members to create an optimal ensemble. But beyond the simple selection of members, it also improves the representation of errors by calibrating the ensemble against actual meteorological data. The terminology is commonly used in the weather forecasting community, from which our technique was first applied in the early 90's. Considering the history of the terminology and the better representation of ensemble statistics, we decided to clarify in the abstract and the introduction for the broader audience but we kept the term "calibration" to preserve the idea of the regularization of our statistics in the later sections.

Abstract, P1, L20: "*Two* ***optimization techniques (i.e., simulated annealing and a genetic algorithm)*** *are used for the selection of the optimal ensemble using the flatness of the rank histograms as the main criterion."*

P4, L7-18: "*In this study, we start with a large multi-physics/multi-analysis ensemble of 45-members presented in Díaz-Isaac et al. (2018) and apply a down-selection or calibration process similar to the one explained in Garaud and Mallet (2011). Two principal features characterize an ensemble: reliability and resolution. The reliability is the probability that a simulation has of matching the frequency of an observed event. The resolution is the ability of the system to predict a specific event. Both features are needed in order to represent model errors accurately.* ***Our main goal is to down-select the large ensemble to generate a calibrated ensemble that will represent the uncertainty of the transport model with respect to meteorological variables of most importance in simulating atmospheric*** CO_2***.*** *These variables are the horizontal mean PBL wind speed and wind direction, and the vertical mixing of surface fluxes, i.e. PBLH. We focus on the criterion that will measure the reliability of the ensemble, i.e. the probability of the ensemble in representing the frequency of events (i.e. the spatio-temporal variability of the atmospheric state).* ***For the down-selection of the ensemble, we will use two different techniques, simulated annealing and a genetic algorithm from now on refer as calibration techniques/process****. In a final step, the ensemble with the optimal reliability will be selected by minimizing the biases in the ensemble mean. We will evaluate which physical parameterizations play important roles in balancing the ensembles and evaluate how well a pure physics ensemble can represent transport uncertainty.*"

REF-C2: Also, in many places, it would be better to replace "errors" with "uncertainty" (i.e. statistical characterization of the unknown errors).

Author-C2: Errors was replaced with uncertainties in some places or the manuscript.

REF-C3: The flatness of the rank histogram is the primary criterion for selecting the ensembles. What doesn't seem to be discussed is the significance of various departures from flatness (as a function of the numbers of bins and the number of samples in the histogram). What fraction of the roughly 8 million possible 6 member ensembles (from the 45 cases) have essentially the same flatness. It is these questions that need to be clarified for understanding of whether the different results from SA vs GA are selecting from different populations of near optimal cases, or whether the differences are pretty much what you would expect from statistically based optimization on a population with a very flat minimum.

Author-C3: The figure below (Figure A4) shows the frequency of the rank histogram scores for each calibration technique, sub-ensembles size and variable (wind speed, wind direction and PBLH). This is based on the sub-ensemble collected at the end of the process, where the rank histogram score and bias are smaller than the original 45-member ensemble. We found more cases in the lower scores for PBLH and in the higher scores for wind speed and wind direction. The figure shows that overall, wind speed controls a significant amount of the optimization because of the high frequency for large scores. However, wind speed doesn't impede the selection of ensembles with a small score for the other variables (PBLH and wind direction). We added this figure to the appendix and make some reference to this point in section 3.2.2:

Section 3.2.2, P13: "The rank histogram scores for all variables are greater than those for one-variable optimization (see Table 4). ***The high-rank histogram scores are associated with the equal weight gave to the three variables for this simultaneous calibration, where wind speed controlled the calibration process. For the calibration of the three variables together, we were not able to produce an ensemble for wind speed with a score smaller than four, this ends up limiting the selection of the calibrated ensemble for the rest of the variables (see Figure A4 in Appendix 1).*** In addition, all these calibrated sub-ensembles have biases …"

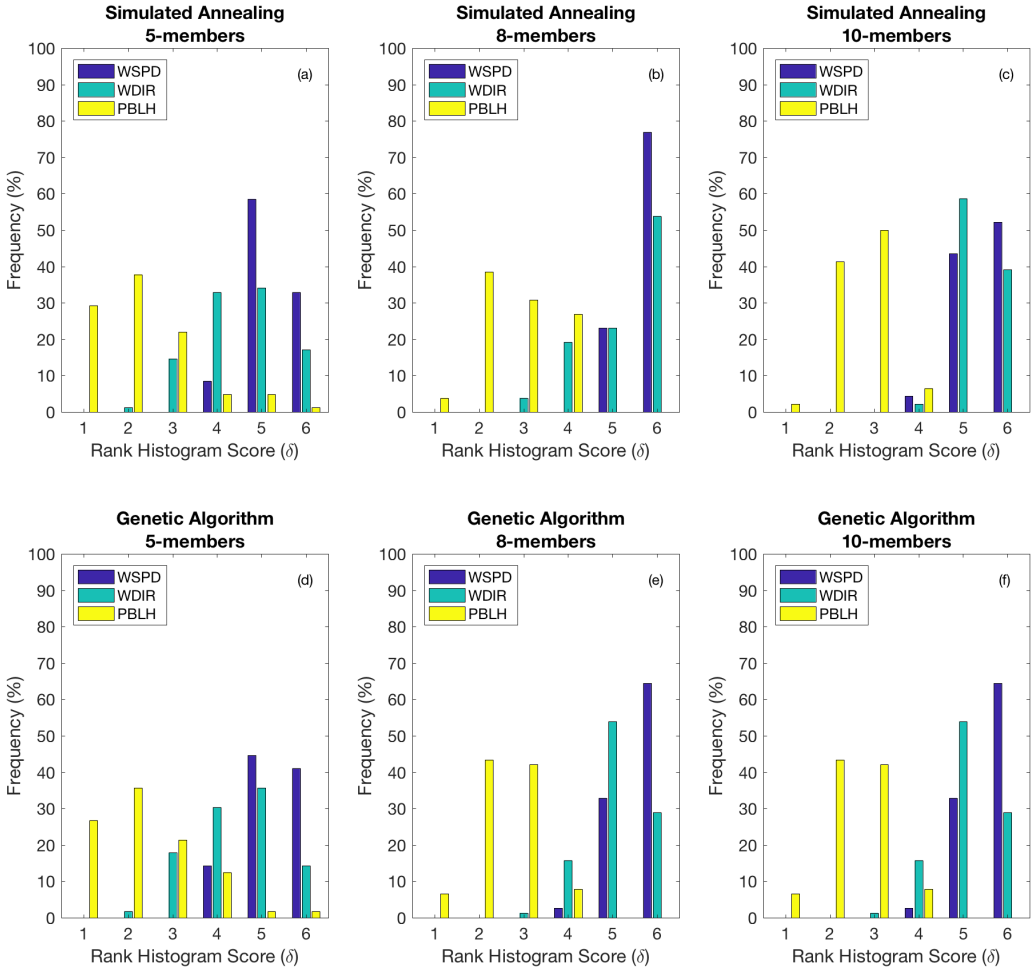

Figure A4. Rank histogram score of calibrated sub-ensembles of different size generated by Simulated Annealing (a-c) and Genetic Algorithm (d-f). Each color bar represents the frequency of that scores for the three different variables wind speed (WSPD), wind direction (WDIR) and PBL height (PBLH).

REF-C4: However, what I don't understand about the SA and GA searches is "why bother". Why not just look at all the cases explicitly (About 3 billion for the 10 member ensembles). For M observations, all you need is a 45 by M table of the p_i (i.e deviation between model and obs). Then generate each sub ensemble in turn. For each case you scan the table and for each of the 45, you count the number where that model is part of your current sub ensemble AND p_i is less than zero. This number tells you which bin to increment. After dealing with all M observations, calculate delta and J As you work through all 3 billion possible sub ensembles, keep track of the best J (and note which ensemble) and any other statistics that you want. This looks like it is well within the capabilities of modern computers. For many purposes, there is no need to store stuff about all 3 billion ensembles, but if you wanted to, you could store all the J values in about the same amount of memory that I have on the sd card in my low-end smart phone.

Author-C4: We tested the brute-force solution at an early stage of the paper and concluded that the size of the sub-ensemble would become very rapidly a limitation. Beyond 10 members, the number of solutions increases very rapidly and requires hours if not days to compute. It is nearly impossible for 20 members or more. We note here that our objective was to use an objective

methodology applicable to any ensemble sizes. We have submitted a second study with a 25-member ensemble which, in this case, means 3,000 billion combinations, hence requiring our Monte Carlo approach.

REF-C5: As a minor point of notation, the ensemble, defined as a set, S, is indicated by upright font when it is introduced (p8, L27) but is shown as a slant font (as used for algebraic variables) as is done on the next line, and in eqn 3 and most later places. The usage should be made consistent. Also, subscripts that are words or abbreviations of words, upright font should be used.
Author-C5: We corrected the S and the subscripts as suggested. Also, we changed *J* for δ to keep everything consistent throughout the article. Please see the next edited part:

"Both techniques generate a sub-ensemble (*S*) of size N. For the first test, we will use these algorithms to choose the combination of members that optimize the score of the reduced ensemble δ (*S*) (i.e., rank histogram score) for each variable. With this evaluation, we determine if each optimization technique yields similar calibrated ensembles, and if the calibrated ensembles are similar among the different meteorological variables. In the second test, we calibrate the ensemble for all three variables simultaneously, where we use the sum of the score squared: $[\delta(S)]^2$:

$$[\delta(S)]^2 = [\delta_{\mathrm{wspd}}(S)]^2 + [\delta_{\mathrm{wdir}}(S)]^2 + [\delta_{\mathrm{pblh}}(S)]^2, \quad (3)$$

to control acceptance of the sub-ensembles. In Eq. (3), $\delta_{\mathrm{wspd}}(S)$, $\delta_{\mathrm{wdir}}(S)$ and $\delta_{\mathrm{pblh}}(S)$ are the scores of the sub-ensemble for PBL wind speed, PBL wind direction and PBLH respectively."