-The difference between the two model configurations is not the model error, but rather one representation of model uncertainty. At 4 km resolution, this uncertainty will pertain systematic differences between the two configurations chosen in this study, and sampling from the data base would mean consistently sample perturbations with the same systematic error. Using the differences between two configurations where one has a known systematic deficiency needs to be better justified, if at all possible.

*We now explain what we mean by model error in a completely revised version of section 2.1.*

-The main short-comings of this study is the computation of the "error" (uncertainty) itself. Here the authors turn off the deep convection parameterization and claim "it is assumed that the turbulence (together with shallow convection) and resolved condensation schemes might compensate for the absence of parameterized convective transport". And they proceed to compute the "error" as the difference in total transport (where one experiment is now missing the convective transport terms). This assumption is highly questionable. Just because there is no parameterized convection contributing to the transport flux of e.g. specific humidity, doesn't mean that there is no convective transport. In the "no parameterized convection" experiment this is now taken care of by the resolved dynamics, and the "compensation" discussed will be seen in the tendency of the dynamics. In fact, the authors do point out in the introduction that studies have shown that turning off the convective parameterization at ∼4 km can lead to unrealistically strong updrafts. What is the scientific justification for systematically adding a positive (or stronger negative) perturbation to the total transport when the convective transport is missing by construction, and is now resolved?

*We have completely revised section 2.2. In fact when switching off the deep convection (DC) scheme, the dynamics starts to handle some of the convective transport that would otherwise be treated by the DC scheme. As is now better explained we study the error that is created in the transport when switching off the DC scheme.*

*Some models run without a DC scheme at resolutions below 5-km. We can also run our model at 4-km resolution without DC scheme. We then demonstrate that an ensemble approach can increase forecast skill.*

-The simulations should be made with non-hydrostatic dynamics, for the dynamics to be able to (have a chance) to realistically simulate vertical motions generated by convection.

*We have taken the necessary time to run the experiments with the non-hydrostatic version (NH) of the model and show the results now in comparison to the hydrostatic ones. First we show explicitly that our hydrostatic model performs better than the NH version for the setup of our experiments (as we expected since the model with the DC scheme was tuned for the hydrostatic setup). It is explained in the new manuscript that DC scheme compensated for some of the vertical transport in the hydrostatic version with respect to the NH one. We perturbed the NH NCP version with the database and show the results sec. 3.2. of this non-hydrostatic version. It essentially leads to the same conclusions. We invite the reviewers to check this with the results of the hydrostatic setup in the first version of the manuscript.*

-The perturbations are applied to the model considered the "target" forecast – which does not use a convective parameterization. Now you systematically introduce a larger

parameterized convective transport in a run with resolved convection. This seem to imply that the scale awareness of the model impose a reduction of the resolved convective transport, such that the improvement that you see relative to the control run (e.g. Figure 9), basically comes from again implicitly "activating" the convective parameterization (by systematically introducing a larger convective transport in the physics parameterizations).

*In fact, when switching off the DC scheme the dynamics takes over some of the transport. This is now discussed in the revised version when discussing Fig. 3.*

-Why the first time-step after turning off the deep convection parameterization is a representative time of the model uncertainty needs to be better justified. The uncertainty due to convection ought to grow as a function of lead time. Figure 3 simply shows total transport with and without convection.

*Indeed we agree that the first version of manuscript was lacking the detail of the theoretical justification and a clear definition of the model error we are studying. We have rewritten sec. 2.1. As a short answer here, we define model errors only after 1 time step, so we do not consider non-linear growth.*

-Another aspect that is rather confusing in the experiment design is why the study is constructed such that the perturbations are applied to the model configuration that has no deep convection at 4 km, when the operational model uses the convective parameterization? Wouldn't it be more desirable to create a perturbation scheme that could be applied to the operational ensemble system at that resolution?

*The underlying question is whether, in an ensemble context, one could represent the subgrid uncertainties related to a parameterization by a stochastic process. When considering the statistics of the model error in the Figures of the PDFs, one can notice some systematic dependencies, from which one could hope to characterize them systematically and to develop a more fundamental stochastic scheme that does not depend on a sampling of a database. This could be tested with some fitting of the model errors in our database. But alternatively we believe such a model-error database could be useful to feed a machine learning algorithm to discover systematic model-errors in the physics parameterization and then use them to perturb the models in an ensemble context.*

-Lastly, the perturbations in the distribution are not applicable to any general model system, but tied to this very particular experiment setup, and thus does not provide a general guidance for development of stochastic parameterizations. What happens if the model is used at 10 km or 1 km? Which configuration is now considered the 'perfect' model?

*We now describe the definition of the model error better in sec. 2.1. We explain better what is meant by "perfect model". It should now be clear that it can be applied generally to any resolution, but the database should then be recomputed. Model errors are certainly resolution dependent.*

Reviewer 2

**General comments**

The use of a hydrostatic model at 4 km and expected to represent convective flows in the tropics realistically is, forgive the word, an aberration. The small improved seen with the use of the stochastic are arguably due to this shortcoming more than anything else. It is well known that although too coarse to represent the details of individual clouds, CRM at few kilometre resolution (up to 10 km in some cases) represent well organized convection in the tropics; the Japanese did their first global CRM simulation at 7 km and got very realistic MJO, CCWs and MCSs. On the other hand it is also well known that the hydrostatic balances messes up gravity waves at scales of 50 km and less and a fortiori convective flows in this range.

*As mentioned above, we have provided outputs for a non-hydrostatic setup in the revised version of the manuscript.*

The way the sampling of the flux errors is done is not very clear. While I am likely confused by their narrative, the choose of the 250hP level as a reference for "sampling the grid column database" is not only not justified by the authors but it is also not accurate. This leaves behind all the convective activity which is associated with shallow clouds of cumulus congests and stratocumulus type.

*Indeed you are right, the methodology of both the error computations and the sampling was not clear in the first version of the manuscript. We have rewritten these section 2.1 and 2.2. providing much more detail that was lacking. As a short reply, we do not sample at 250 hPa but we sample entire vertical profiles. This is now made clear in the revised version.*

Tropical convective systems are known to involve a rather diverse population of cloud types and one needs to account for all of them in order to represent the life-cycle of organized convection. According to the authors, the whole argument for choosing to simple a flux-error database instead of the more or less established Stochastically Perturbed Parameterization Tendency (SPPT) is rooted from the fact that the error fluxes associated with different variables are only weakly correlated (if they are at all).

*No, the errors profiles are weakly correlated to the model profiles of the total transport. But you have a good point: the text could be written more clearly. We have adapted it. We now write: "Large correlation coefficients between the <u>model</u> transport flux and its error would suggest a linear relationship ..."*

However, the way they do the sampling while it does assume such correlation it makes it systematic since they sample the grid columns and not the different fluxes independently as illustrated in Figure 8.

*Our excuses but we do not understand the last sentence.*

**Specific comments**

Lines 20-25 of page: This paragraph is misleading when first reading it the following question came to my mind: "Something is not quite right. How can one compare fluxes between two different models that do not necessarily go through the same integral

curve in the state space?" It is only after I got to page 4 that I found out that the authors are doing the right thing by comparing flux deviation after only the first step. This needs to be stated before hand so not to confuse the reader.

*Very correct, sec. 2.1 and 2.2 were badly written. As mentioned above we have reworked them thoroughly.*

Line 15, Page 4: "Therefore, the retrieved model error should rather be seen as a lower bound on the error made in the representation of the physical process." This can be interpreted as that the authors are trying to do better than the reference? This may not be possible since the direction of error can not be quantified in such a large dimensional state space!

*Also in this case, this part has been completely rewritten.*

Page 4, line 27: The use of a hydrostatic model at 4km resolution needs caution–while I doubt that it can be justified, the authors are requested to provides a few words warning their readers that this is not at all realistic! This is intact a serious flaw in this study. A quick look at the Gerard et al. reference reveals indeed that the cumulus scheme on which this study is based tries to represent non hydrostatic effects (their Eqn. 5), thus it is not surprising if the deficiencies in the NPC model are have more to do with the use of hydrostatic model other than anything else.

*As is now explained in the revised manuscript, we can also run the model with a non-hydrostatic dynamical core (NH). We have rerun our tests wit the NH dynamical core. We present scores in sec. 2.1 comparing both of them. We have implemented the perturbation in both hydrostatic and non-hydrostatic version. In the revised manuscript we provide the results with the non-hydrostatic version. Which lead, by the way, to the same conclusions as the ones for the hydrostatic model (which can be verified with respect the previous version of our manuscript).*

Line 25, page 5: "This database is not only useful to investigate the statistics of the model error due to deep convection parameterization (Sect. 2.3), but it will also be the basis for a stochastic perturbation scheme that can be applied in an ensemble prediction system (Sect. 3)." Rephrase or delete the whole sentence. It adds nothing to the paper it can only confuse your readers. Isn't the later statement the main objective of the study?

*Indeed, it is now deleted.*

Figure 2: This figure can be clearer. It took me maybe 5 minutes of staring at it before I could make a clear sense of it. The caption could be used to explain the labels and the color coding.

*We now explain the figure in great detail in the full text in sec. 2.2 to make the whole method more clear.*

Page 6, line 3: Aren't 72 evaluations too few given there is a high level of correlation in space and time because of the nature of organized convection?

*We found an improvement with this. Of course, with more evaluations we might expect even more improvements. This is a matter of computational resources. As mentioned to review 1, we plan to proceed investigating our model-error database. This study can be seen as a first feasibility/sanity test.*

Figure 3: What does the label error in red stand for? It isn't clear at all. The red dots are hardly visible and they don't constitute and error but their difference does. Maybe draw a red line segment between the two red markers to indicate the error.

*It is the first time step after switching off the deep-convection scheme. This is now better explained in the revised version of the manuscript. We think the confusion comes from the lack of description in sec. 2.1 and 2.2 of the previous version.*

At- page 7, line 9: The discussion in these two paragraphs and Fig 3 seems to be included in order to make the final statement that "Therefore, the total transport flux difference one time step after the switch can be considered as a representative measurement of the error in the transport flux as defined in Eq. (1)." 1) This is empirical observation has no scientific value as such. 2) The model error as defined in (1) is only valid when evaluated at first step because the states of the two simulations change in subsequent steps.

*Again, we define and compute the model error after 1 time step, see the new description in sec. 2.1 and 2.2.*

Page 8, line 5: This is not a surprise because the model needs to conserve the water budget.

*Indeed, this is a sanity check. But we prefer to not include this in the new manuscript.*

Page 11, line 5: This is not a surprise at all because the model needs to conserve the water budget.

*Idem as your remark above.*

In reply to your three points:

- Page 13, lines 13-14: The way the sampling is done is not at all clear.
- Figure 5 has three distributions, which one is actually sampled.
- Figure 8, has six fluxes how the two are reconciled? Are you sampled the distributions in Fig.5 or the "grid columns data base"? If it is the later how are you doing it? Is it uniformly over all grid columns? Also Conditioning on the basic state would been more appropriate if one wants to genuinely emulate the cumulus scheme. Nonetheless the "success" of the completely random sample in reproducing the results implies that the cumulus parameterization is perhaps not sensitive enough to the environment, which may be problematic.

*We have rewritten the paragraph starting with "The vertical and inter-variable correlation are preserved by organizing the flux-error profiles per grid column in the ...". In fact the perturbations are multivariate as we now explain in this paragraph.*

Page 14, lines 1-2: Why are you doing this? Aren't the cases with zero or weak updraft part of the physics of the problem? This is clearly biased and it is not at all justified. It undermines the role of shallow cumulus and cumulus congests clouds since your distributions in Fig. 5 are based on 250hPa errors.

*Since we only study model errors originating from the deep-convection scheme we will exclude these point from the model-error database. We write this explicitly in the paper. Also we think your confusion comes from a poor description of the definition of the model error in sec 2.1. of the previous version of the manuscript.*

Page 14, lines 8-21: So you are using a convection trigger. Are the two criteria enforced simultaneously or are you using one at a time? Why these particular choices? How do they compare to what the original cumulus scheme does?

*No it is not a trigger. They are two criteria to test whether the gridpoint has deep convection or not and to see whether it goes into the model-error database. This should now be clear with the sentence mention in reply to your point above.*

Figure 9, caption: "Lead times where the ensemble mean RMSE is significantly lower than the NCP control RMSE at the 95 % confidence level are indicated with a filled circle." This is not clear that this is actually true. Maybe showing the absolute errors instead would be more clearer. In any case the difference between the compared errors is probably very small. What is the actual gain really is?

*It is difference in error. So negative values mean improvements.*

Page 15, line 5: This may have something to do with perhaps the fact that you are sampling the flux errors at 250 hPa in Fig. 5.

*As mentioned above we are sampling vertical profiles not errors at 250 hPa.*

Page 15, lines 7&9: CPâĂŤ>NCP

*In fact this is not shown. But it is true, the stochastic scheme beats the parameterization in the hydrostatic case. We have taken this sentence out of the manuscript since we are now showing the non-hydrostatic results.*

Page 15, lines 12-13: When and where the error in the reference configuration was it defined? You can't really tell since you are not comparing to anything else but the CP run. Please delete this sentence.

*It is now defined in sec. 2.1.*

Page 17, lines 4-7: This is in contradiction with the claim made upfront that a stochastic parameterization would increase the spread by accounting for model error.

*But it increases the spread. We only compare it here to the spread coming from the IC/LBC perturbations.*

Page 17, lines 11-13: This applies to any ad hoc and non-physically based CP.

*Indeed. But it is also the case here.*

Page 18, line 15: Where are you looking? Do you mean MOCON and OMEGA?

*Indeed, the text is wrong here. We now write: "Considering the spread in Fig. (15), for horizontal wind, ..." It confirms what ones expects. Models without deep-convection parameterization tend to produce grid point storms, and this may wrongly generated extra spread in an ensemble.*

Page 22, lines 5-6: What does this mean? Are the two models evaluated and compared elsewhere? If so please provide the reference and eventually say for which case study it was done. It makes a huge difference if that was done for tropical or non tropical convection site. Otherwise simply

delete this sentence. It simply says that in the gray zone the role of a CP is unclear whether it is beneficial or detrimental and this is already known for many years.

*Yes we now compare, CP to NCP in both hydrostatic and non-hydrostatic setups in sec. 2.1.You are right it is not clear whether it is beneficial or not. We deleted the sentence.*

Page 22, line 7: This isn't true.

*We are severe here for ourselves. We refer here to the confidence interval, the key word is "significant" in the sentence. We adapted the sentence. We now write: "There is a neutral to positive impact in skill for the MOCON ensemble (albeit inside the significance confidence intervals)."*

Figure 16 actually shows the opposite. The NCP ensemble is better than the MOCON ensemble during the first 9 hours.

*Indeed, we present it as it is.*

Page 24, lines 4-5: " but for many variables it even outperforms the ensemble system with the deep convection scheme switched on." Where is this shown?

*It is not shown. It is nevertheless true. It was a sentence from a previous draft of the manuscript.*

*Given the new figure in Fig. 1 and the detailed description in sec. 2.1. This is the take-home message of the paper. One can characterize model error due to shortcoming of the parameterization and then use it to perturb the model in an EPS sense to find that it can restore some predictability in a probabilistic sense.*

Page 24, line 10: spell out EPS

*This is done.*