RESPONSE TO REVIEW #2 OF GMD MANUSCRIPT:

"Introducing the Probabilistic Earth-System Model: Examining The Impact of Stochasticity in EC-Earth v3.2"

We thank the reviewer (hereby referred to as Reviewer #2) for their comments, which have helped us improve the clarity and quality of our manuscript.

1) **Reviewer #2:** *"Title: the title mentions the "Probabilistic Earth-System Model" while indeed all work has been done entirely with the atmosphere component of EC-Earth only. In my eyes the title this suggests more than what the manuscript delivers and to avoid any too far reaching expectation I'd therefore suggest to modify the title to something more adequate."*

Our response: The reviewer makes a fair point. We have now edited the title to "Progress Towards a Probabilistic Earth System Model: Examining The Impact of Stochasticity in EC-Earth v3.2", reflecting the paper as a natural step from the vision outlined in Palmer (2012), "Towards the probabilistic Earth system simulator". This should be more in line with the actual content of the paper.

2) **Reviewer #2:** *"Section 2.1, 1st paragraph: there is no need to go into details about EC-Earth's ocean component or coupler because they are irrelevant in this work."*

Our response: We have removed these details.

3) **Reviewer #2:** *"Section 2.2, eq 1: what is "the i'th physics parameterization scheme"? You mention the schemes explicitly towards the end of Sec 3.1, why don't you list them here already?"*

Our response: We have now explicitly identified the different physics schemes in section 2.2.

4) **Reviewer #2:** *"Section 2.2, eq 1: are the tendencies for all variables perturbed by the same r? Or are there different perturbations for the different variables?"*

Our response: All variables are perturbed in the same way: the tendency vector P has, as its entries, the tendencies for temperature, humidity, and wind fields. It is this vector which is perturbed, implying that all variables are perturbed similarly. We have slightly rephrased the explanation of what P is to emphasize that it contains all these tendencies to clarify this point.

5) **Reviewer #2:** *"Section 2.2 , eq 1: are the pertubations constant in time, or do they vary from one timestep to the next?"*

Our response: The perturbations evolve over time as AR(1) processes. We have rephrased the presentation slightly in response to this point, emphasizing both that the perturbation happens at each timestep, and that the perturbation r evolves in time.

6) **Reviewer #2:** *"Section 2.2, end of 1st paragraph: you say "perturbation is limited between 0 and 2" but shouldn't that be between -1 and 1? Even with r=-1 we still get that P_hat has the same sign as P, or?"*

Our response: This was incorrectly phrased on our side: the reviewer is entirely correct that r itself is clipped between -1 and 1: this implies that 1+r is clipped between 0 and 2, which is why this range was referenced in the submitted manuscript. We have revised the script accordingly.

7) **Reviewer #2:** *"Section 3.1, p.6 l.13: why do you use ERA-40 or ERA interim for the evaluation? These re-analysis datasets belong to the same model "family" as EC-Earth and may therefore share common biases, in particular where the re-analysis are largely a model product not constrained by observations. I would prefer the evaluation to be done against a different re-analysis dataset, but that may be too much to ask for at this stage. In any case, it would be good if the similarity between EC-Earth and the model used for the re-analysis would be clearly stated."*

Our response: The reviewer makes a good point. Experience suggests this is unlikely to be an issue for surface variables such as surface temperature which are strongly constrained by existing observations and therefore do not differ much across reanalysis datasets. However, certain less strongly constrained variables, such as total cloud cover (considered in the paper) may be more affected by this, so we agree it is important to point out.

We have now added a disclaimer addressing this point in section 3.1 covering this point, and we hope the reviewer will find this sufficient.

8) **Reviewer #2:** *"Section 3.4: how good is it to compare only the last 10 (5) years from the last ensemble member against observations? Is that short time period really sufficient to confirm or reject the fitness of EC-Earth? I don't understand why you don't average 10-20 year from all ensemble members (start dates) and compare against the corresponding time period of the re-analysis (see also comment above about choice of re-analysis data)."*

Our response: The reviewer makes a very reasonable point. As implied in the manuscript, we restricted our assessment to this time period because the default CTRL version of the model was tuned for this particular time period. The reason for restricting attention to this shorter time period is a decision made by the EC-Earth Consortium as a whole, but I believe it is to avoid the possibility of 'fitting' the model's climate sensitivity. If one tries to match the energy budget over a long time period, one is effectively trying to fit the specific temperature growth, i.e. the climate sensitivity over the historical period. Fitting therefore to a very recent time period, where there is good satellite data for robust observational estimates, is therefore a pragmatic choice for model tuning.

One key conclusion drawn from this paper is that the model will in general need to be retuned after the addition of stochastic schemes, and that for this reason, comparisons of a scheme's performance against the CTRL model must always be considered with respect to the model's tuning procedure. In particular, because EC-Earth is tuned according to the historical time periods energy budget, the performance of a stochastic scheme on the energy budget over the same period will almost always look like a degradation. This is why, when comparing to observations, we only explicitly considered the short time period in question.

It is true that we could nevertheless have considered the evolution of the energy budget over the full simulation period 1960-2000. However, this essentially amounts (as explained above) to assessing the impact of the schemes on the model's climate sensitivity. This was not a topic we wanted to expand upon in this manuscript, because it is the topic of an independent manuscript currently under review (Strommen et al 2019, now cited in the revised manuscript). We have added a caveat to this effect in Section 3.4.

9) **Reviewer #2:** *"Section 4,5,6: when you talk about cloud water do you mean the gridpoint value of the in-cloud value? I would suspect it's the gridpoint value, and if that's the case then the reduction in cloud cover would imply an even larger increase in the in-cloud water content with all the consequences for optical thickness and cloud microphysics. Good if you could clarify on this point."*

Our response: Cloud liquid water refers to the vertical integral of liquid water contained within clouds in a single gridpoint column: we have added this descriptor to the revised manuscript in response to the reviewers point.

The point about reduced cloud cover further `exacerbating' the increased cloud water content is a good one. We have included a line to this effect in the revised manuscript, section 4.1.

10) **Reviewer #2:** *"Section 4.2,5.2,6.2 and Fig 8: I don't see a point in making a timeseries of the energy budget, the interesting aspect for a climate model is how well it simulates the average flux and its variability compared to observations (re-analysis). For that reason it would make more sense to average the biases in the fluxes and present them as a table or barplot similar to how you did in Figs 1 and2."*

Our response: While we agree with the reviewer that the most important aspect is the model's ability to capture the average flux and variability compared to observations, the timeseries plot does capture some relevant further information. Because the differences are effectively constant in time, this implies that the impact is not independent of the initial ocean state. Furthermore, one can infer that the model adjustment is extremely rapid, with the stochastic schemes not causing any systematic drift; because of the potentially slow response of the land-scheme, such drift cannot be ruled out a priori even in the absence of ocean coupling. Therefore, we feel there is value in keeping this as a timeseries plot.

We do however accept that the above reasoning was not adequately included in the submitted manuscript. We have added a brief comment to this effect in sections 4.2, 5.2 and 6.2 (Energy Budget impact sections for the three schemes).

11) **Reviewer #2:** *"Section 7: what is the motivation for selecting the Hadley circulation and the QBO as test cases for the stochasticity? Why not NAO or PNA as other prominent modes of atmospheric variability?"*

Our response: In response to a comment by Reviewer #1, we have added motivation for these choices in the revised manuscript: see the introduction to Section 7.

12) **Reviewer #2:** *"Section 7.1 Fig 9: It is not easy to easy to distinguish the different colors of the dots (I am slightly color blind) and you should consider presenting the data in a different way, e.g. by using different symbols for the different members/time periods."*

Our response: We have edited Figure 9 to make the points be larger and use different symbols; we also added a legend to aid interpretation. The colourscale has been chosen due specifically to its readability for colour blind people: if the reviewer still finds the plot tricky to parse, then we would be very happy to edit it further !

13) **Reviewer #2:** *"Section 7.2: what is the reason for evaluating QBO only for the last ensemble member? Why not using the results from all start dates?"*

Our response: We were not clear enough in the presentation of this section. To be clearer, the estimate of the QBO period for the simulations are done using all 5 ensemble members: the period shown is the average across these 5 estimates. For Figure 11, showing a pressure-time contour plot, we only show this for the last ensemble member because the degraded QBO looks extremely similar across all 5 members. Therefore, we opted to simply show the last member as an example, against the more recent ERA-Interim reanalysis data.

We have added some clarifying remarks to this effect in Section 7.2, in response to this point. In addition, we have included errorbars in the QBO period plot, to show the spread across the 5 estimates. This illuminates the results further and immediately suggests that the full ensemble is used for the computation.

14) **Reviewer #2:** *"Section 8, 1st sentence: you cannot blame the absence of a more process-oriented analysis on the lack of available data because you have designed the experiment and its output."*

Our response: Yes, you are of course right. The point is rather that we made an explicit choice to construct experiments that would illuminate changes in the long-term climate, and therefore did not construct an experimental protocol that was suitable for robust assessment of the rapid response of the model. We have rephrased the introduction to Section 8 to clarify this.

We further point out there, that experiments aimed at illuminating the rapid response of the model to SPPT were carried out in Strommen et al (2019). These in fact appear to confirm the hypothesis made in the present paper, as they show that the very rapid response is to cloud liquid water and evaporation.

15) **Reviewer #2:** *"Section 8.1, p.14 l.7: it's not clear to me how an increased in-cloud liquid water content could steepen the near-surface humidity gradient. Could you explain better what you mean?"*

Our response: This was poorly phrased on our part. We have rephrased this part of section 8.1 in response to questions from Reviewer #1, and it is hopefully more illuminating now. For your convenience, this paragraph now reads:

"With both SPPT and ISPPT, the dominant impact on the energy budget is increased evaporation. In the IFS, the amount of evaporation at a gridpoint depends primarily on the surface wind-speeds and the extent to which the specific humidity at the surface gridpoint differs from the saturation humidity (a function of surface temperature). While wind-speeds do increase by about 1.4\% on average with ISPPT, the mean wind-speeds are unchanged with SPPT, with a tiny increase of only 0.06\%. Given that the increase in evaporation of both SPPT and ISPPT are of the same order of magnitude, this suggests changes in humidity are a key factor. Because SST's are held fixed, such changes will be, to first order, driven by changes in the water content of the atmosphere as opposed to temperature changes at the surface. One possibility is that the increase in cloud liquid water is depleting the near-surface humidity, causing more favourable conditions for evaporation. The fact that both cloud liquid water and surface wind-speeds increase more with ISPPT could then explain why this impact is amplified in those experiments. Another possibility is that the first order impact is on convection in the tropics, which may be activated more frequently with SPPT/ISPPT. This could lead to a drying of the boundary layer, promoting more evaporation in response."

16) **Reviewer #2:** *"Section 8.2, p.15 l.13: what do you mean with runoff being a key driver of atmosphere interaction? Isn't runoff simply the difference between P-E and the amount of water absorbed by soil? It's a residual, not a driver, or?"*

Our response: It would indeed be more correct to say that soil moisture is the key conduit between the land and the atmosphere, through its impact on evaporative cooling and latent heat transfer. The point we are trying to make in the manuscript is that the primary impact of the LAND scheme is to affect runoff, and the changes in runoff lead to changes in soil moisture and, hence, the atmosphere as a whole.

Note that because the scheme is perturbing parameters in the soil equations themselves, it is possible for the scheme to change runoff *directly*. Note also that while *surface* runoff is effectively a residual, *subsurface* runoff in the land surface model is generated through a free drainage condition, which is dependent on soil hydrology parameters.

17) **Reviewer #2:** *"Section 8.2, p.15 l.22: why do you call runoff a tuning parameter in the LAND case? Runoff is an important diagnostics of the model that can be used to tune the model, but runoff itself isn't a tuning parameter."*

Our response: There seems to be a misunderstanding here. Our statement is "Tuning parameters for EC-Earth **include constants that regulate** […] **runoff** in the LAND scheme case" (bolded for emphasis). In other words, we absolutely agree that runoff itself isn't a parameter that can be tuned, but other actual tuning parameters have a strong impact on the behaviour of runoff in the model, and these parameters *are* tuned.

We have therefore left this statement as is in the revised manuscript.

18) **Reviewer #2:** *"Code availability: EC-Earth is licensed and not openly available, it's not sufficient for any presumptive user to request access. I would suggest you check the guidelines of GMD that regulate code availability and re-phrase this section."*

Our response: Yes, this should have been made clearer. We did state that access was available upon requesting permission from the consortium, but should have been clearer that this is due to explicit licensing issues attached to the IFS component. We have now rephrased this.

19) **Reviewer #2:** *“p.2 l.12: "However" seems to be inappropriate in this sentence.”*

Our response: We agree, and have rephrased this to “Most modern climate models also …”.

20) **Reviewer #2:** *“p.15 l.15: shouldn’t it rather be "...none of the schemes _is_ able..."?”*

Our response: This particular grammatical construction is one of those funny English loose ones, where I believe one can get away with using either ‘is’ or ‘are’. However, when the meaning is `none of them’, one typically uses the plural ‘are’.