We would like to thank the reviewers for reading and reviewing the work, which we think has lead to substantial improvements. In the following, we copy the reviewer comments by reviewer/item (*brown*, *blue*, *orange italic texts*) and provide our replies in black, with changes in the manuscript inserted in magenta and new analysis (appendix material) in purple.

1 Replies to reviewer 1

"Convective physics are a vital part of climate models but lead to lots of uncertainty and biases. As convective processes tend to occur on spatial scales smaller than can be resolved by most models, they must be parameterized. These parameterizations are vital to understand as they control lots of model results, but are enormously complex and have effects that are hard to isolate.
In this manuscript, the authors place data from a full GCM into single column models to see how the physics model interacts with the column when it is isolated from the dynamics. They find that all the physics models tend to decrease CAPE through boundary layer drying, and that there are free troposphere changes associated with precipitation.
The manuscript is interesting and the framework for analyzing model biases is valid. However, the analysis could be much more thorough and the results are not very general. Additionally, the manuscript has serious grammatical errors and inappropriately cites "In preparation" references."

We thank the reviewer for carefully interpreting our work and are glad that they find the manuscript interesting and will reply to the specific comments individually in the following.

"Specific Comments

- *Extremes vs. Typical Points: Figure 2 shows that the largest values of CAPE from the models is larger than those of ICON. Presumably the columns where this is occurring behave very differently than the boundary layer drying occurring "on average" – what is going on there? Relatedly, figure 5 (left) shows the mean and the IQR. For most points, the mean is outside the IQR, likely due to extremes dominating the mean. The IQR is not useful as an uncertainty for the mean, it is appropriate for the median only. Also this suggests more broadly that extremes behave very differently from normal situations, so averages may not be the best metric. Overall, given the non-linearity and asymmetric distributions inherent in convection, it is necessary to be more careful with regards to taking the mean and considering what is typical/likely"*

We agree that the behaviour of the extremes of the CAPE and precipitation distributions is also of major interest. To understand this behaviour, we have carried out some further analysis of the occurrences of mixed layer CAPE exceeding 7000 J/kg at 6 hours lead time across our models.
The summary of the further investigation is that EOFs analysis is an appropriate method to understand their variability and individual EOF components may be held responsible. Hence, we believe that this is rather relevant for this reply and the following summary should be provided as supplementary material or appendix to our study. Here, we highlight the key results:
The mixed layer CAPE may reach extreme values in excess of 7000 J/kg in situations where the three-hourly forcing tendency is rather extreme. Thereby, there is significant commonality between IFS and GFS. For instance, when IFS exceeds 7000 J/kg at +6hrs, its conditional mean CAPE is 7846 J/kg (87 samples, about one in 100.000). For the same samples, the GFS CAPE then averages 8563 J/kg. However, while for RAP and ARPEGE, the conditional means are elevated, they are instead in the range of 5000-7000 J/kg. The RAP suite has the most

distinct pattern of extreme CAPE occurrence, whereby far more samples exceed 7000 J/kg than in each of the other three suites (1278, one in about 7000). The conditional mean of the other suites is between 1000 and 4500 J/kg over those samples.
The tendency patterns which dominate the high CAPE sample are also relatively common across IFS, ARPEGE and GFS. Samples signify destabilisation of the mixed layer by the advection of humidity (IFS) or both humidity and temperature (GFS). The advective humidity tendencies in the mixed layer are at centered at about 7 standard deviations from the mean in IFS and undercompensated by a physics response at 2.5 standard deviations (opposite sign). Similarly, GFS physics undercompensates the strong dynamics tendencies for this subset at only 1.6 standard deviations (again, opposing the large-scale forcing). A similar pattern can be identified for GFS, but there both temperature and humidity tendencies undercompensate the advective forcing of humidity (centered at 5 σ) and temperature (6 σ) in the mixed layer. ARPEGE shows something intermediate between GFS and IFS.
However, the extremes in RAP are driven predominantly by free-tropospheric tendencies. Again, excessive forcing is responsible. However, the balance looks very different: free-tropospheric dynamics tendencies in humidity center at about 13 σ from their mean in the excessive CAPE subset, but the opposing physics compensates at 16 σ. The temperature tendencies are much better in balance. At the same time, the humidity forcing in the mixed layer is also very anomalous (though less than for extreme CAPEs in IFS, GFS and ARPEGE) and undercompensated by native RAP physics.
Along with the corresponding table shown below, we choose to add the above as appendix material (Appendix A).

Selection criteria:
MLCAPE > 7000j/kg
_+6h forecast

n_std_conditional_mean
_relative_to_model_mean

Highlights:

Major balance issues

Minor balance issues

Of forcing-physics pairs.

IFS	sample size: 87			
	IFS	GFS	ARPEGE	RAP
delta_t_dynamics_FT	0.13	0.09	-0.41	0.09
delta_t_dynamics_ML	6.52	6.58	4.66	6.58
delta_q_dynamics_FT	-0.93	-0.90	-1.86	-0.90
delta_q_dynamics_ML	7.15	7.19	6.03	7.19
delta_t_physics_FT	1.07	1.11	0.89	0.97
delta_t_physics_ML	-0.48	0.67	1.84	0.20
delta_q_physics_FT	-0.45	-0.45	-0.03	0.36
delta_q_physics_ML	-2.54	-1.63	-5.98	-4.37
Mean CAPE	7846	8563	5144	6211

ARPEGE	sample size: 256			
	IFS	GFS	ARPEGE	RAP
delta_t_dynamics_FT	-0.61	-0.64	-0.81	-0.64
delta_t_dynamics_ML	2.75	2.78	2.29	2.78
delta_q_dynamics_FT	0.81	0.84	0.20	0.84
delta_q_dynamics_ML	6.54	6.52	7.20	6.52
delta_t_physics_FT	2.03	1.90	1.82	1.77
delta_t_physics_ML	-0.16	0.84	1.27	0.55
delta_q_physics_FT	-1.22	-1.26	-1.38	-0.75
delta_q_physics_ML	-3.94	-2.28	-2.25	-3.26
Mean CAPE	4489	5759	7705	5201

GFS	sample size: 440			
	IFS	GFS	ARPEGE	RAP
delta_t_dynamics_FT	-1.24	-1.27	-1.66	-1.27
delta_t_dynamics_ML	6.30	6.34	4.61	6.34
delta_q_dynamics_FT	0.11	0.13	-0.57	0.13
delta_q_dynamics_ML	5.04	5.03	4.47	5.03
delta_t_physics_FT	2.35	2.10	1.99	2.10
delta_t_physics_ML	-0.89	1.37	1.87	0.45
delta_q_physics_FT	-1.74	-1.74	-1.20	-1.10
delta_q_physics_ML	-2.76	-0.25	-5.54	-3.37
Mean CAPE	5191	8083	4027	5370

RAP	Sample size: 1278			
	IFS	GFS	ARPEGE	RAP
delta_t_dynamics_FT	-15.69	-15.64	-15.91	-15.64
delta_t_dynamics_ML	0.88	1.04	1.30	1.04
delta_q_dynamics_FT	12.62	12.58	12.58	12.58
delta_q_dynamics_ML	4.87	4.79	4.07	4.79
delta_t_physics_FT	16.40	16.32	16.04	16.31
delta_t_physics_ML	-0.08	0.07	1.08	1.96
delta_q_physics_FT	-16.40	-16.67	-14.61	-16.40
delta_q_physics_ML	-5.46	-2.42	-1.20	-0.37
Mean CAPE	1133	4089	4233	8048

Consequently, we would also like to reflect on the second aspect of the reviewer comment. We agree that non-linearity (and the associated vast range of spatial and temporal scales) could have an important impact on our experiment, its analysis and synthesis. The philosophy behind the manuscript is that the EOF analysis provides the final insight in this regard, while the initial impression could be that CAPE destruction and precipitation diagnose our spin-up in isolation, continued analysis (Section 3.3-3.4) indicates that mixed-layer drying could explain most or all of our spin-up. Finally, the conclusion is that mixed-layer drying is indeed important, but precipitation is mostly a response to forcing in the free troposphere and various modes of boundary layer/low-level dynamics also shape our dataset, where some include imbalanced mixed-layer humidity tendencies. However, the exact patterns are quite diverse among the models and not so simple to appreciate.

Nevertheless, they account for CAPE destruction and associate much more weakly with precipitation variability than is suggested initially. This sequence is a good analogue for multiple scientific levels understanding of what is apparently a simple problem when you start analysing an experiment. The first EOF shows that while the distribution has an extensive tail, the free-tropospheric tendencies can be decomposed into one EOF and nevertheless conserve its dominant variability. However, a multi-variate perspective (of two to three EOFs per model) is required to fully digest the boundary layer/mixed layer behaviour emerging in our tendencies, and spin-up is prevalent. Our limitation to four mixed layer tendencies (with affect DCAPE) imply that the requirement of two to three EOFs for the mixed-layer to conserve its dominant variability is a proxy for the strong non-linearity. Further variables would be needed to fully map out plausible non-linearity, which requires alternative techniques. However, the free tropospheric variability is much more digestible in our SCMs, despite the complexity of convection. It is (perhaps surprisingly) strongly dominated by a very linear mode. Four of our tendencies represent the free tropospheric variability, and in addition, precipitation variability (and CAPE changes partially) reflect free tropospheric variation: well over 40 of our variance and, quite possibly, up to 60% of our EOF dataset. Out of this, 48% is related to EOF1 in each model - an estimated fraction of 80 to 100% of the the free-tropspheric variability included in our EOF analysis. This, together with the above analysis, leads us to the conclusion that linear analysis of the free troposphere indicates not only surprisingly much similarity among our SCMs but it also explains much of the activity of model physics in our domain.

- *"Generality: While the number of columns analyzed in this study sounds like a lot, they are all from the tropical Indian ocean within one month of each other and one model. They represent two interesting cases (ITCZ and subtropics), but miss out on all other seasons, elsewhere in the world, other SST regimes, and any other models. The authors somewhat acknowledge this, but the study would be much improved with a few other examples (other models, other regions, other times of year).*
  One straightforward way to generalize would be to use idealized profiles. Rather than running a full model, create synthetic, simplified, profiles for different situations and put those in the models. This could also make it clear what the dependencies are; if we create a synthetic profile with even more humidity in the lower troposphere, does CAPE decrease by more? This could simplify/supplement the conclusions from the EOF section"

We do agree with the reviewer that it would be of very high interest to extend our MUMIP experiments to the extratropics. Since the first experiment and phase has taken a few years, this is admittedly beyond our current scope within this study. However, a more extensive inclusion of extratropical domains is of interest for future MUMIP experiments. We recommend so in the revised discussion and have thought about sensible hypotheses (revision: from line 640):
Our analysis indicates that variability among different physics suites in the free troposphere has large similarity, consistently with Groot et al., 2026 (JEMS), and future analysis of future MUMIP experiments in higher latitude domains could target similar diagnosis to ours at higher latitudes. Thereby, a key hypothesis would be to test whether the clean separation of an EOF representing the precipitating part of the hydrological cycle also exists at these latitudes, in connection with free-tropospheric tendencies.

- *"Writing issues: The citations to "in preparation" papers limit this study's reproducibility and have little authority if those papers cannot be seen. Additionally, there are grammatical errors throughout, here are the first two as an illustration:*

- *Abstract, lines 13–15: "a imbalance between these terms in the mixed-layer correlates with the CAPE change at least one of them"*
- *Line 17: "may have serve various objectives"*

We acknowledge that works with many references in preparation are difficult to assess and will keep it to a minimum. Nevertheless, works on full methodological overview of MUMIP with further sensitivity experiments are in advanced stages of preparation and can be relevant to specific statements, but not critical. Finally, the work Groot et al., 2026b has been published in Journal of the European Meteorological Society and it can be appropriately referenced.
We have corrected typos and errors in grammar/writing in the revised version, including the two you highlighted.

2 Replies to reviewer 2

"General comments: The authors investigated spin-up behaviour in four single-column models within the MUMIP framework, with a particular focus on convective diagnostics. The same external forcing from the ICON-DYAMOND experiment is applied to all four SCMs, and the evolution of precipitation and CAPE over the Indian ocean domain are examined by looking at correlation and EOFs. The study offers two potentially novel contributions, one is a quantification of spin-up variability across different models, and the finding that CAPE change is more strongly correlated with mixed-layer drying, which is generally expected to correlate with precipitation in the mid latitude.
The underlying scientific question is timely and interesting, and the MUMIP dataset is genuinely novel. I believe the study contains publishable scientific findings. However, I found the manuscript difficult to follow, making it challenging to clearly identify the main scientific message in its current form.
For this reason, I would encourage the author to seek additional internal feedback on the manuscript structure and readability prior to resubmission. I have listed my major concerns below, focusing primarily on issues of presentation and organisation. I would welcome the opportunity to evaluate the scientific contents of the work in greater depth once the manuscript has been revised and its central conclusions are more clearly articulated."

We thank the reviewer very much for the constructive attitude (and detailed commenting in the following) as well as for recognising our intents.

"Specific comments:

1. *Missing no spin-up setting*
   To separate spin-up effects from the evolution of the background state, it would be valuable to include a baseline experiment that is not expected to exhibit spin-up effects (e.g., ICON SCM forced by ICON output). The authors instead include the ICON-DYAMOND output as a benchmark, which I consider a reasonable substitute if used consistently throughout the analysis. However, this benchmark is only shown in Figs. 1 and 2 and is absent from the subsequent analyses. At the same time, the manuscript continues to discuss the ICON benchmark results (e.g. L260, L273, L394), including the statement "the ICON benchmark tendencies would be negligible". It was not always clear to me whether this statement is supported by analyses not shown in the manuscript or is inferred from the experimental design. I therefore encourage the

authors to ensure that all discussion points are clearly linked to presented results and that the ICON benchmark is treated consistently throughout the paper."

Unfortunately, it is not possible to construct a configuration that lacks spin-up.
If one would compare the ICON DYAMOND with any other ICON configuration, spin-up will emerge. The coarse-grid ICON weather prediction model can be shown to differ just as much from the benchmark in terms of convection-dynamics feedback loops as our SCMs — e.g. see Groot et al., 2024 (WCD, Copernicus) for an ICON-ICON comparison which shows that the feedback from the physics to the dynamics depends heavily on grid spacing near the convective grey zone. All in all, in our opinion all model intercomparisons and initialised/assimilated experiments have to inherently suffer from some kind of spin-up, and all we can do is use clever methods to diagnose it (which we aim to present here).
We include a brief explanation of why a spin-up-free SCM configuration is not feasible in the revision as given below (from line 133 in the revision):
Unfortunately, it is impossible to provide a plausible spin-up-free reference experiment as a control case. This is because of differences between convection-permitting models and coarser grid models (Groot et al., 2024, WCD, Groot et al., 2026, JEMS): for example, comparison of multiple coarse-grid ICON ensembles with a convection-permitting ensemble indicate consistent differences in convection-dynamics feedbacks for various settings, including an ensemble without convection parameterisation. These results largely generalise to multi-model physics suites in MUMIP (Groot et al., 2026). Furthermore, each ICON ensembles would have a different humidity climatology at the same location (Wing et al., 2020, JAMES) than coarse-grid or LES versions of ICON. This combination of constraints excludes possibility of producing a spin-up-free SCM dataset. Besides this problem, providing much higher resolution (temporal and) spatial SCM forcing is computationally unfeasible.

2. *"Reliance on preprint and in preparation.*
   The manuscript currently depends heavily on preprint by Groot et al. (2026b) and in-preparation studies. Including additional methodological and scientific context within the manuscript itself will make the manuscript more self-contained and would broaden accessibility."

We would like to refer to the final reply item to comments of reviewer 1 (also on "in preparation" works).
Finally, we will further address the comments of the reviewer on structure and methodology where these are detailed (i.e., in the following parts of the review, item 4).

3. *"Missing concluding remark. The authors state their aim in this paper as "Our overall aim is to quantify spin-up and physically meaningful variability with the statistical analyses. Thus, we here identify which spin-up plays an important role in the MUMIP and whether physics tendency patterns can be generalised across multiple models." These objectives provide a useful framework for the study. However, I found it difficult to identify where they are explicitly revisited and answered in the later sections of the manuscript. The Summary section restates individual analysis results, but does not clearly synthesise them into conclusions. Consequently,*

I found it challenging to assess the extent to which the analyses support the conclusions corresponding to the objectives implied by the Introduction. I encourage the authors to include a dedicated Conclusions section that explicitly revisits the original objectives and summarises the main findings in terms of the posed questions."

We thank the reviewer for pointing this out and add a conclusion section in the revised version (revision: lines 693-706):

Diagnostics of moist physics and convective instability reveal spin-up in the MUMIP Indian Ocean experiment as signified by mixed-layer CAPE and accumulated precipitation rate evolution. Empirical orthogonal function (EOF) analysis of tendencies of four physics suites shows that their mean free-tropospheric response to commonly prescribed large-scale dynamics within our domain can largely be summarised into a common EOF. This EOF is associated with precipitation, but it is only weakly associated with changes in our convective instability metric: change rate of mixed-layer CAPE. The result indicates that dominant tendency patterns arising from complicated physics suites are generalisable.
The mixed-layer tendencies project more strongly onto the change rate of CAPE in most (if not all) models. The boundary layer tendencies project onto multiple EOF patterns. These tendencies are not generally in balance in any model and, therefore seem to project onto the transient CAPE evolution.
Specific humidity (re-)distribution is a key issue in weather forecasting and climate models, as also found in other studies. Therefore, it would be ideal if humidity (re-)distribution could converge across all contemporary simulation tools without their loss of accuracy.

4. *"Structural and stylistic problems. The manuscript contains conflations of materials that would conventionally belong in different sections and paragraphs. For example, the paragraph starting at L48 in Introduction appears more appropriate for the method section. Similarly, the EOF methodology would be easier for readers to follow if introduced in the method section rather than in the result section. Another example appears in Section 3.3. The first three paragraphs seem introducing the scientific background, which might fit more naturally in introduction. These are followed by correlation analysis results and then by a discussion of the Hadley-cell dynamics underlying the results. Those are all within one subsection of Results. Similar mixture repeatedly appears throughout the paper. I therefore recommend a substantial restructuring of the manuscript so that readers can more easily follow the scientific narratives.*
   I also encourage the author to have a thorough grammar, and consistency review. Examples include inconsistent line-break styles (e.g., L66, L317, and L323) and very short paragraphs (e.g., L426 and L477) that interrupt the flow of the text. Addressing these issues would ensure that readers can focus on the paper's inherently valuable scientific messages."

We have made some changes to improve the readability and structure. The paragraph of L48 of the original submission in the intro has partially been moved to the methods section and the remaining sentences are now merged with the following paragraph of the introduction. The reason to follow this structure was to familiarise the reader early on firstly with the physics suites as entities rather than with individual parameterisation (and their closures) and secondly the concept behind our multi-model experiment, which is key to this work.
The small review on convective closure, quasi-equilibrium and CAPE (change rate) in section 3.3 is relevant for the local discussion at that point, which distinguishes from review material on entire

physics suites (introduction) rather than the individual *closure* perspective. To balance the need to separate these types of material,keeping related topics together, we have denoted this as '3.3.1 background material' before diving into the analysis of the newly introduced 3.3.2 on our results. We moved methodological parts regarding our EOF analysis into the methods (and modified them). We perform a grammar and consistency check and hope that it resolves the main issues pointed out by the reviewer (together with the added conclusion; the revised EOF methodology starts at line 172).

We apply the classical EOF analysis (Lorenz, 1956, available via MIT.edu) and investigate a dataset of SCM tendencies (see table 1) in Section 3.5. Thereby, we normalise variables with respect to their standard deviation.
Physics and dynamics tendencies of our dataset are available, and their residual causes model state changes in humidity and temperature:

$$\delta = D + P \tag{1}$$

Where D is the explicitly resolved dynamics tendency in three-dimensional models as prescribed in our experiment, P is the net physics tendency obtained as sum of parameterised tendencies and δ is the residual net tendency, which modulates the state variables. Furthermore, we assume the following about our large multi-model SCM tendency dataset:

- the long-term mean of the sum of physics and dynamics tendencies should go to zero in a stationary climate simulation with (near-)stationary forcing ($\overline{\delta}$ -> 0, where the bar denotes temporal averaging and some regional averaging)
- the long-term mean of the change rate of CAPE should go to zero under (near-)stationary forcing
- any net physics tendency P is expected to be statistically compensated by the opposite dynamical tendencies D. Theoretically, this is valid in the long run and any model in its native climate should closely follow this assumption on much shorter time scales.

Statistical imbalances in tendencies over large areas will cause temperature (humidity trends), hence, climate trends. These signals have rates corresponding at most to the seasonal cycle in multi-day free-running climate simulations (i.e., $\overline{\delta}$ at K/month) or even lower in multi-annual climate simulations (generally $\overline{\delta}$ < 1K/decade). We can safely neglect these in the ICON benchmark, which implies assuming 0 tendencies is safe, since net model tendencies are on the order of K/day.

5. *"Minor specific comments: 1. Fig. 5: I recommend to include a plot for DCAPE."*

Thank you for the recommendation - indeed, initially, this seemed like a very promising way of presenting our work: we have considered adding such a plot from the beginning and carefully thought about whether it has added value for the overall narrative. However, in the end, a panel with DCAPE turned out to be unnecessary for the story line and it would not provide interesting information beyond the existing Figures in our eyes. Since all physics suites are initiated to reproduce ICON initial conditions, the CAPE PDF in Figure 2 of other physics suites looks initially indistinguishable from that of ICON. The change of CAPE is directly identified from the translation of the curve of each suite at 6 hours leadtime. While the PDF is visualised on a logarithmic axis (which leads to difficulty when interpreting distribution differences), major shifts can easily be identified and set the scene for the remainder of the study. Figure 5 displays interquartile ranges to

further help the reader with a non-logarithmic axis. Although quantile crossing (i.e., each respective quantile at initial time is represented by a different column at 6 hours lead time than at initial time) asserts that Figure 5 does not directly represent CAPE change, the broad tendency envelope is also clearly visible here.
Because DCAPE is by definition zero or undefined at initial time, lead time dependent information cannot be provided. Instead, below, we show the time series of DCAPE across all the models and the temporal mean ratio between IFS DCAPE and CAPE, each at 6hrs lead time.

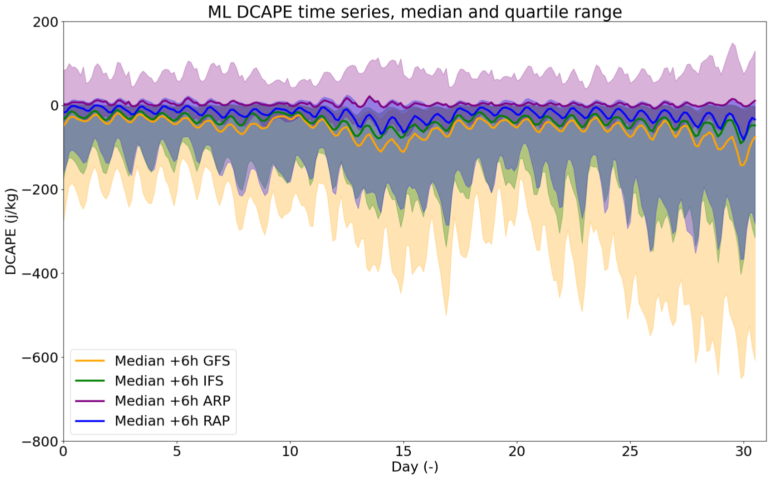

Figure 1: Time series of DCAPE in each model; ARPEGE is abbreviated as ARP. Here, the median and interquartile range are based on the entire domain. The statistical structure of these tendencies is nicely reflected in the equatorial statistics, shown in Figure 5 of our study (acknowledging that this band covers about a quarter of our domain).

"2. L161: It is stated that "The precipitation rate is broadly similar for five configurations. That means no systematic bias between the convection-permitting benchmark and our physics suites appears, although this might be expected." I do not agree with this statement. They seem similar because of the logarithmic scale. Precipitation rate is the most known biased variable among models and it is one of the largest reasons why we develop expensive convection-permitting models. If a SCM forced by coarse external forcing can produce the same precipitation as finer-resolution models, we don't need to invest the development. I found that showing a raw PDF with a logarithmic axis and concluding no bias is deliberately misleading."

Our apologies, we have overstated the intended information in the original submission. We reflect on a few reasons to produce convection-permitting simulations and/or use stochastic physics separately in Groot et al, 2026 (JEMS). We have revised the text to increase its accuracy and indicate that interested readers may find more information in the JEMS work (from line 216, as found below).
Figure 1 shows the PDF of precipitation rate of all output columns across four model physics suites and the benchmark, the convection-permitting ICON, when represented at a common grid. Millions

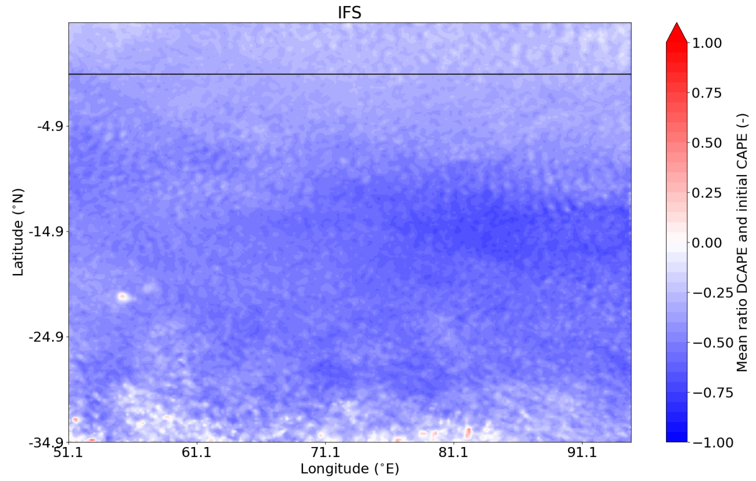

Figure 2: Mean ratio between DCAPE and CAPE in IFS at 6 hours lead time. Our equatorial band shows little statistical variation in the temporally averaged DCAPE-CAPE ratio.

of columns receive negligible precipitation accumulation (see also Groot et al, 2026 (JEMS), for a discussion of trace precipitation). On our logarithmic axis, we find a rapidly decaying PDF for low intensities (less than 20 mm/6hrs) in ICON, which looks like an exponential decay on a log-axis. From about 20 mm/6hrs, precipitation rate decays in an approximately loglinear fashion, with virtually no event above 90-100mm/6hrs.
The precipitation rate PDFs are broadly similar for five configurations. No systematic bias between the convection-permitting benchmark and our physics suites appears (when accounting for the range of multi-model spread; see also Groot et al., 2026 (JEMS)), although this might be expected. For an extensive discussion of the joint precipitation PDFs of physics suites and our benchmark, including their small conditional biases, we refer to Groot et al, 2026 (JEMS). This work also suggests a coupling of precipitation to the dynamical forcing, which we also address further in the current work.

3 Replies to reviewer 3

"The manuscript describes a series of experiments where various physics schemes were coupled to a single dynamical simulation in an offline setting. While this is an interesting experiment, I had severe difficulties following the logic of the manuscript. In addition to lacking clarity, the author's interpretation of their EOF analysis is questionable and needs to be revised. overall comments

1. *It was not very clear to me why exactly such an experiment (coupling physics to a common, non-native model) is relevant in the first place. Perhaps you can give some more motivation, what exactly do we hope to learn about either these physics suites or about tropical convection in the real world? "*and *"I find it hard to draw any general conclusions from this paper. What have we learned beyond "different physics suites behave differently"? Perhaps a section*

5 "Conclusions" could be added?"

We thank the reviewer for their comments. We have revised the motivation. We want to understand the differences between models (physics suites), identify spin-up and address generalisability. This has been revised at the end of the introduction, where the main interest of this work is defined. Furthermore, we hope that by accommodating for specific comments of reviewer 3 and the suggestions of reviewer 2 the interpretation difficulties are tackled and resolved (also when complemented with a section dedicated to the EOF explanation). Finally, by adding a conclusion (see replies to reviewer 2) we hope that the scope of this paper has been clarified (revised lines 78-80).
Our overall first aim is to assess whether we can quantify spin-up adjustment. Subsequently, we aim to decompose tendency structures of our experiment into model-dependent spin-up adjustment components and statistically balanced tendency components, which are hopefully comparable, with significant overlap among physics suites.

2. *"There is a lot of text describing relatively basic diagrams like pdfs or profiles in verbose detail. Consider shortening section 3 wherever possible. The discussion and interpretation of the EOF analysis is confused and probably misleading in some ways (see comments below). In particular 1) the entries in the PCs are neither correlations nor do they represent the physical magnitude of some tendency and 2) physical interpretation of the higher EOF modes is usually impossible because of the orthogonality constraints. "*

Unfortunately, a very detailed PDF discussion is required to appreciate and introduce the opposing changes between the main modes of the distributions (in both CAPE, LNB), the change in variability and the opposite tendency of the extremes. This assures that we warm the reader up for some multi-variate effects later on in the EOF analysis. Furthermore, we have identified that the various reviewers still have questions regarding the various Figures. However, we cannot discuss every detail and hope we have made reasonable choices on what to include and exclude with the revised submission.
Finally, we firstly reply to the specific questions about the EOF analysis in the respective items below (mainly item 29 and a few others). Secondly, by including Section 2.3 with the detailed assumptions, we hope that the lines along which the EOF is interpreted are clear after revision.

"specific comments

3. *"l.13f unclear grammar ("at least one of them"? who or what "are explained"?) please rephrase"*

We have rephrased (revised lines 13-14).
EOFs of each suite indicate that an imbalance between these terms in the mixed-layer is connected with the CAPE change in at least one of these two

4. *"l.17 "have" or "serve"?"*

"Serve"; we have rephrased.

5. *"l.25 "persist in datasets with increasing sample sizes" unclear what you mean, which sample size, the length of the simulation?""*

We have rephrased the above (revised lines 25-28):
(...) if they persist in datasets with increasing sample sizes (upon statistical convergence), i.e., for increasingly large ensembles, for many forecasting experiments in weather prediction, or in long free-running climate simulations

6. *"l.29 unclear grammar, I feel like the verb is missing?"*

We have rephrased for clarity to (revised line 29):
While model tendencies are often utilised to investigate a single model, specific processes, individual parameterisations and/or variables (...), the sources of tendency variance can also be utilised for targeted uncertainty assessment (...).

7. *l.50f unclear what the brackets ("observations, validation of MUMIP") refer to"*

Rephrased to (revised line 51):
This forcing is derived from the simulation of the Icosahedral Nonhydrostatic model (ICON) performed at 2.5 km grid spacing within the first phase of the DYAMOND project (Stevens et al., 2019, PEPS) (equivalent to "observations", which represents an analogue of a data assimilation procedure in numerical weather prediction). Four physics suites forced in single-column versions of models from three different groups have contributed to our dataset.

8. *"l.60 unclear what you mean by "more accurate". Clearly a free-running model has no correlation with the actual weather observed at any time (no skill), but that's probably not what you mean?; l.63 "supposes more rigorous uncertainty sampling" what is that supposed to mean?"*

We have rephrased to improve the clarity (revised lines 56-60).
Thereby, we focus on weather time scales as opposed to solely on climate time scales in many of the other "MIPs". Numerical models initialised for weather prediction are known to be more accurate when assessed for coherency of their evolution of atmospheric variability with the real world than typical free-running climate simulations (e.g., Bauer et al., 2015, Nature). This is due to advanced initialisation using data assimilation, and to a lesser extent, due to finer grids in horizontal and vertical direction, tailored perturbation techniques and smaller time steps.

9. *"l.70 unclear what "them" refers to, the subject of the previous sentence is singular."*

"Them" has been rephrased to (revised line 68): distributions of each aforementioned diagnostic.

10. *"l.110 unclear in what way a subset is extracted and what it means to "mimic ICON dynamically""*

We have rephrased this to (revised lines 114-116):
From all columns we extract a major subset for which the dynamical forcing has been validated correctly across all SCMs (with correlation and variance analysis) (...)

11. *"l. 127 which is it, 30 or 60? or were both used?"*

It is 30 mins for IFS and 60 for other physics suites (see revised text).

12. *"l.136f what exactly does it mean that you "mixed a layer of 500m depth"? does that mean the parcel to be lifted has the average conditions of that layer? from where is the ascent initiated?"*

Lifted parcels should be lifted from the centre location of the mixture. We have added this (revised line 156):
The mixed layer potential temperature and specific humidity is weighted by mass, with the bulk lifted from the centre of this layer.

13. *"Figure 1: lines are hard to distinguish, consider (a) trying a log x-axis, (b) showing the frequency bias with respect to ICON (ratio between the four physics suites and ICON);*
Figure 1: unclear what you mean by "ICON & initial conditions", are those two separate things?
Figure 1 (last comment): why was the y axis cut off at 10? for some applications (especially weather forecasting), the tail of the distribution might be the most interesting part.
l.155 and elsewhere: figure 1 (also 2) does not show pdfs (densities) but raw counts. That won't make a difference to the figure unless the sample sizes are different for some reason."

We see problems with the suggestions of the reviewer: a logarithmic x-axis would only compress the tails further. Secondly, the ratio of two logarithmic values would be undefined once one model has no samples in a bin, which is a problem.
We have increased the display size of Figure 1 in the revision. The "initial conditions" text was a mistake.
The reason to cut off at 10 is because a sample size of 10 means that the PDF shows a probability of less than 1 in 1.000.000 in a bin. This is already in the far end of the tail. The tail of the distribution already starts at order 100s to perhaps 1000s of samples (near 30 mm/6hrs, 120mm/day). We therefore do not consider the tail excluded. As the sample sizes shrink to 1-3, the error bars will increase and this may disproportionally distract from the remainder of the PDF (since we have chosen a logarithmic axis).
Finally, we have indicated below Figure 1 that each of our histograms do indeed translate to PDFs - given the prescribed constant bin sizes.

14. *"l.178 what is "clearly stable", is that defined thing or just a turn of phrase?"*

We have rephrased for clarity (revised lines 239-240):
The vertical integral signified by CAPE represents the total amount of available potential energy when parcel of low-level air is forced upward until it freely convects up to an inversion layer, such as a capping inversion or the tropopause.

15. *"l.183 these two peaks (at least the first one) are not so easy to see in the figure.*
Figure 2: naively, I would expect that LNB height and CAPE are connected and at least looking at the tails of the distribution ist seems to me that the two panels actually show similar things with a non-linearly compressed x-axis. Did you look at the joint distribution of CAPE and LNB? Are they closely linked?"

We have made the Figure larger. The peaks could be diagnosed by finding a minimum in the curve of each of the physics suites. This should be fairly doable for most physics suites, but we agree that it may be more tricky for GFS. From the existence of a minimum, it follows that there are at least two peaks.
Yes, LNB and CAPE will be correlated, but very large to extreme CAPE does not necessarily lead to high LNB: LNB variability is more strongly related to variations in tropopause height, while CAPE is more strongly connected with steeper mid-level lapse rates and low-level humidity. Still, it turns out there is quite good correlation (values like 0.7, decreasing with lead time for RAP, for instance).

16. *"l.233 where have we seen anything about the tendencies of cape? figure 2 only showed the pdf of cape itself, not of its tendencies? "*

We have emphasised in the revised first paragraph of the methods that each physics suite reproduces the initial conditions of ICON (in revised lines 89-91):
Initial conditions are borrowed from a common benchmark of the DYAMOND project (Stevens et al., 2019) and, accordingly, reproduced accurately by each suite.
This, together with the discussion about the PDFs indicating CAPE changes should tell the reader explicitly that CAPE tendencies are associated with the general distribution changes (and differences among physics suites) in CAPE.

17. *"Figure 3 and 4 why not include the ICON reference here? "*

The choice was intentionally made to exclude ICON after Figure 2. The remainder of the storyline leads from spin-up diagnosis to area-mean correlations to the EOF analysis. ICON does not have a clearly defined lead time comparable to the physics suites: it is near-free-running at 10 to 40 days lead time. Therefore, the comparability ends with the spin-up. ICON will spatially show similar patterns, as GFS and RAP do. However, we have also omitted them to channelise the information. We also channelise the information by focusing on the equatorial band for the EOF analysis and CAPE spin-up panel (Figure 5).

18. *"Figure 4 why is cape not shown in terms of w_{max} here? wouldn't that make it easier to interpret and compare with the pdfs for before? "*

We agree that two choices may be defensible for equally appropriate reasons. Since we argued for w_{max} in the PDFs to improve the visualisation, we have a choice between sticking to this, or choosing for the more intuitive classic CAPE. Since CAPE maps can widely be found at the web, we chose for CAPE maps.

19. *"l.258 all the curves look pretty much linear near the end of the 6h, where to you see a reduction in rain rate? "*

Yes, this is what we meant. Precipitation rate reduces rapidly from the first to the second hour, and perhaps a bit from the second to third. We have therefore clarified our statement by labelling this statement with a specific time interval.

20. *"Figure 5 left: the areas are very hard to distinguish, for most lines we cannot see whether the mean is within the quartile-range (see below). Is this the distribution over the whole domain or just the equatorial band? If the former, wouldn't it make more sense to show the same region as in the right panels, so we could compare them?*
    Figure 5 left and l.264f: since precipitation has a skewed distribution, the mean can lie outside the interquartile range (see last timestep for GFS). If you want to use that range to gauge the uncertainty, maybe it makes more sense to show the median instead of the mean? The text discusses this behavior but doesn't mention the simple explanation that the distributions are non-gaussian."

To further increase the clarity, dotted lines are added to delineate the range of shading in the revised Figure 5. We also increase the size of the Figure for improved readability. We have explicitly indicated that the left panel is about the full domain. The left panel illustrates that the full experimental domain has model-dependent spin-up. The right panel subsets for the equatorial region because the CAPE is an order of magnitude smaller in the CAPE, which reduces the clarity of the spin-up if diagnosed from the panel. Furthermore, it connects well with the EOF analysis later. We could have added the precipitation in the equatorial band separately, but this would not provide much information complementary to the left panel (with similar patterns and since most of the precipitation falls in the ITCZ region). Finally, given the many panels that we could have shown, we have been considerate with reasonable choices. We hope that this provides the reader with the essentials.
We agree that the skewed distribution of precipitation is often interesting. In Figure 5, the interquartile ranges on the left provides typical values of precipitation that are generally exceeded in the ITCZ region, which consists of about a quarter of the domain (not a coincidence). On the other hand, the lower quartile is representative of the driest parts of the subtropics in the experiment. The mean curve is dominated by the quarter of the domain associated with the ITCZ, but smaller contributions to the mean are produced in the subtropics. The quartiles are meaningful because they are broadly indicative of spatial variance, and exact quantile choices are not super important when complemented with Figure 1 (and 2 for CAPE). Of the distributions are non-Gaussian and they are explicitly shown in Figure 1, so this should be very clear.

21. *"l.272f and elsewhere, you discuss what the ICON benchmark "would" look like. Why not just include it in the figures?"*

This is because we cannot add lead times of 10 to 40 days to Figure 5 with an existing margin between 0 and 6hrs leadtime (which is explicitly emphasised in the revised version).

22. *"l.307 I was very confused for a minute because DCAPE usually refers to the energy available for convective downdrafts (https://glossary.ametsoc.org/wiki/dcape/), but from the units in figure 6 I'm guessing you mean the time derivative of regular CAPE? Please chose a different abbreviation. "*

Yes, unfortunately, there are two definitions of "DCAPE" in literature, namely downdraft CAPE and DCAPE/Dt with the omission of "/Dt". We have chosen for δCAPE in the revision and updated Table 1 accordingly, where it was defined.

23. *"figure 6 isn't the most striking observation that ARPEGE has only positive tendencies and all other models have only negative tendencies? I don't understand that at all, how can CAPE*

tendencies have unique sign? Is that because the models always adjust the initial profile to their own climate? Also the correlations with precipitation have opposite signs, how can we interpret that?"

It is a striking feature, but the same holds for persistently negative tendencies. It means that the equatorial band consistently "wants" higher CAPE in ARPEGE and lower CAPE in other models when initialised from the ICON benchmark. The sign should not be unique. We have added Section 2.3 to provide extra guidance when interpreting the tendencies along this line. We will come back to this later, because it is more relevant to points made below.

24. *"l.312 f: do I understand correctly that you correlate spatial mean dCAPE/dt with spatial mean precipitation? If so, why? Wouldn't it make a lot more sense to correlate them for each gridpoint individually? As it is, you might have precipitation cells feeding on CAPE in some regions while CAPE is still being generated elsewhere and the two effects average out?"*

Yes, but because we first start from the perspective of existing literature and revise along these lines (Buschow, 2024, QJRMS). Therefore, we start with the analysis of 3.3. Later, we indeed move to the column-by-column analysis because the spatial mean is a limiting factor for interpretations. At first, we can therefore suggest a simple explanation about equatorial CAPE: mixed-layer drying is connected strongly with CAPE. With the bias analysis of Section 3.4 this is provides a nice access point to 3.5 - the EOF analysis.

25. *"figure 7 if you want to define what you mean by delta, do it somewhere in the text or just refer to delta T here since there is only one variable with a delta in this figure."*

Yes, this quantity is introduced and defined in Section 2.3 of the revision.

26. *"l. 337f just because area mean tendency does not coincide with area mean CAPE change, we cannot conclude that other processes must be at work: the tendencies can be non-zero at every gridpoint, but CAPE tendencies can remain zero in many locations, depending on the overall temperature and moisture profile (for example if there is no LFC)."*

We agree with the reviewer that 0 tendencies in CAPE are often associated with substantial tendencies in other variables and remove the content from interpretation of the results.

27. *"Figure 8: unclear what you mean by "covers 6 out of 8 initialisations".*
    Figure 8: Solid and dashed lines are hard to distinguish, and why is there no uncertainty bound for hour 1?"

We have created both Figures, but with uncertainty lines for 1hr lead time, no valuable information is added (similar uncertainties as 6 hours lead time). The information about 1 hour lead time is just to convey the idea that the spin-up dominates the first hour and reduces later on. We have also specified which initialisations are included, which is all complete initialisations that were available for analysis of this study. The 18 UTC initialisation is incomplete and has thus omitted and the 21 UTC initialisation has not been validated correctly for ARPEGE, which means it was also omitted. Finally, we have revised Figure 8 to portrait format in order to make it much more readable.

28. *"Section 3.5 please clarify whether these EOFs were estimated from area means of (as I assume) from individual gridbox values. If the latter, especially don't understand why area means were used in figures 6 and 7 because now we can't relate those results to this EOF analysis. "*

Indeed, the column-by-column analysis was used. This is specified in both the title of the subsection and at the start of the subsection in the revision. The choice for area-means in Section 3.3 is explained under item 23 of the reply to this reviewer.

29. *"l.410f the formulation "we assume the following" makes it sound like these are assumptions that somehow go into the EOF analysis (are require for it to work or something). Is that the case or are these more like your "expectations"? Either way I have no idea how assumptions 1 and 2 could be related to an EOF analysis since they refer to the mean values whereas EOFs depend only on the covariance matrix. In particular l.417 makes no sense to me, the EOFs are literally invariant under changes of the mean. "*

Yes, they are invariant under changes of the mean. The underlying assumptions are clarified and have moved to the new Section 2.3. Because physics and dynamics tendencies sum up to a residual net tendency delta, there are the following options:

- Large samples of dynamics and physics are correlated with each other, but not with their residual delta. This subset consists of two sub-options:
  1. The intercept between physics and dynamics tendencies is zero. A numerical model will be stable.
  2. The intercept between the physics and dynamics tendencies is not zero. Unless the large sample intercept goes to zero and the physics and dynamics tendencies are correlated, but virtually uncorrelated with net tendencies, the model will drift (spin-up) or the model climate will change (essentially both drift and this use of "climate change" could be interpreted as the same thing).
- Large samples of dynamics and physics tendencies are correlated with their residual delta, but not nearly compensating each other. The model climate will also drift.

Because of the above, physics and dynamics tendencies must be of the same order of magnitude, even scale closely and any imbalance will show up either in the EOF analysis (difference amplitude between physics and dynamics tendencies) or because of non-zero intercept in the relation between physics and dynamics tendencies (near-perfect balance in EOF analysis but spin-up in the model climate).
We have interpreted and explained the EOF analysis strategy along these lines as follows in the revision (revised lines 172-211):
We apply the classical EOF analysis (Lorenz, 1956, available via MIT.edu) and investigate a dataset of SCM tendencies (see table 1) in Section 3.5. Thereby, we normalise variables with respect to their standard deviation.
Physics and dynamics tendencies of our dataset are available, and their residual causes model state changes in humidity and temperature:

$$\delta = D + P \tag{2}$$

Where D is the explicitly resolved dynamics tendency in three-dimensional models as prescribed in our experiment, P is the net physics tendency obtained as sum of parameterised tendencies and δ is the residual net tendency, which modulates the state variables. Furthermore, we assume the following about our large multi-model SCM tendency dataset:

- the long-term mean of the sum of physics and dynamics tendencies should go to zero in a stationary climate simulation with (near-)stationary forcing ($\overline{\delta}$ -¿ 0, where the bar denotes temporal averaging and some regional averaging)
- the long-term mean of the change rate of CAPE should go to zero under (near-)stationary forcing
- any net physics tendency P is expected to be statistically compensated by the opposite dynamical tendencies D. Theoretically, this is valid in the long run and any model in its native climate should closely follow this assumption on much shorter time scales.

Statistical imbalances in tendencies over large areas will cause temperature (humidity trends), hence, climate trends. These signals have rates corresponding at most to the seasonal cycle in multi-day free-running climate simulations (i.e., $\overline{\delta}$ at K/month) or even lower in multi-annual climate simulations (generally $\overline{\delta}$ < 1K/decade). We can safely neglect these in the ICON benchmark, which implies assuming 0 tendencies is safe, since net model tendencies are on the order of K/day.
Over long time scales statistical imbalances in $\overline{\delta}$ are undesirable and indicate model issues, which we assume are absent among our models. Furthermore, it may be noted that the balance of at short time scales is utilised in classical quasi-equilibrium type closures for convection parameterisations for CAPE specifically (see Arakawa and Schubert, 1974, JAS, Yano and Plant, 2012, RoG). If our models do not closely follow statistical balance on short time scales and across our dataset at all, a few reasonable statistical hypothesis may be opted, namely:

1. The residual δ is almost random and, therefore barely correlated with D and P. This is possible only if the statistics of D and P cancel out exactly, because they form the only contributions to δ (Equation 2).
2. The correlation pattern between D and P is out of balance: δ correlates with D or P and transient spin-up will occur. Posterior statistical balance of P and D is required for a stable model. Therefore, spin-up must co-vary with the tendencies themselves.
3. There is considerable spin-up, whereby specific humidity and/or temperature drift away from the initialised state ($\overline{\delta}$ non-zero) because of a systematic bias between $-D$ and P. However, co-variances between D and P could still be in balance.

In summary, statistical balance between D and P is a necessary condition for a stable model climate with small δ, but the statistical balance alone is insufficient for the stability.
By combining EOF analysis with tendency analysis and diagnosis of key quantities (convective instability, precipitation) we can learn about the balancedness of each physics suite within our experiment, commonality of balance across the four models and we could estimate whether imbalance option 1 or 2 are responsible for a specific spin-up pattern. We investigate the balancedness of EOFs in Section 3.5.
As we perform EOF analysis, a limitation is that we assume linearity. Finally, multiple EOF components may co-vary with a pair of physics-dynamics counterparts, but since the EOFs will be orthogonal, unconditional balances should apply to individual EOFs.

30. *"l.438: I know what you mean but the values of the principal components are not correlations. If you want to know whether these things are correlated you have to go back to the matrix."*

Despite our assumptions indicating that it is indeed associated with correlation in this particular case, we choose for "not associated" in the revision (and rephrase similar statements elsewhere).

31. *"l.444f: similarly, the entries of the principal components cannot be directly interpreted as the physical magnitude of those terms. In particular, you standardised all variables, which means that the physics tendencies could theoretically be an order of magnitude larger than the dynamics and still have the same amplitude values here. Even if you had not standardised, the EOFs still say nothing about the overall mean tendencies: if one of those tendencies has a near-constant positive value it will contribute next to nothing to the EOFs but might be a big part of the overall balance. "*

We agree with the reviewer that this is normally the case. However, given our specific assumptions, the order of magnitude difference amplitude is ruled out (see also item 29 of the reply to reviewer 3).

32. *"l.451 this is one of the main things I would actually conclude from this EOF analysis, that ML and FT are relatively independent wrt these tendencies for most of your models. The other thing would be that precipitation shows up only in EOF1 and thus has little to do with the ML variables. "*

We agree that this is an important point and have mentioned this separation in the conclusions. However, because of the assumptions of Section 2.3, there is considerable potential to conclude more from the EOF analysis (see also item 29 of this reply and the next item).

34. *"sec 3.5.2 I did not read this in detail because the discussion of higher order EOFs, apart from the most general characteristics, is often futile: the shape of EOF N is dictated by the requirement of orthogonality to all N-1 higher EOFs, interpreting them in terms of physical consequences is therefore dangerous at best. "*

This is true. However, in our particular case and using our assumptions of Section 2.3 (item 29 again), we can identify paired variables in EOFs 2 and 3 which are occasionally mildly out of balance and interpretability appears extends down to EOF3-4, because EOFs 2, 3 and 4 describe variance structures with substantial commonality as does EOF1. This is why we argue that providing an interpretation of these is justified in our particular case. Nevertheless, small perturbations in EOF2 influence EOF3 and the balance analysis is definitely somewhat less robust in EOFs 3 and 4 than in 1 and 2.

35. *"l.507 unclear what you mean by "increases", increases compared to what?"*

We have specified in the revision that we meant "from initialisation to 6 hours lead time".

36. *"l. 541f if the second point is virtual temperature, it is not "separated" from humidity. l.554f how do you know that Tv adjustment "will be dominated by T adjustment", do you have a reference for this claim?"*

We agree that true separation is not possible. In the thought experiment, we linearise to understand small perturbations and utilise the relationship of McDonald, 1963 (BAMS). The impact of humidity perturbations on virtual temperature is small: if we would go from 25 g/kg to 0 g/kg (with a molecular mass of 29 for dry air and 18 for water vapor), this translates to about 1% density, or 3K near 295-305K (tropics). However, humidity perturbations are typically two orders of magnitude smaller and most of the time at least one order of magnitude smaller (see also Figure 8). Accordingly, Typical virtual temperature perturbations due to humidity are 0.03K or so, and they would rarely reach 0.5K. At the same time, we can use the McDonald relationship and the temperature changes to estimate the pre-condensation change. This is what we do. The result is described in our text.

37. *"l.572 I have no idea what you mean here or whether this was "argued" in section 3.4. What does "scales well with most samples" mean? "*

We have clarified this text in the revision, because it was unclear. What we have described is that the equatorial domain has a slightly stronger low-level drying than the domain mean, because it has higher humidity. Some analysis (not shown) indicated that humidity tendency in the mixed layer scales very roughly with the humidity itself across much of the domain and the SCMs. Hence, the following text appears in the revision (revised line 617):
In Section 3.4 we have indicated that the mean adjustment of humidity scales very broadly with the humidity (...).

38. *"594 again I'm afraid you cant interpret the fact that EOF1 is most similar among models in such a way: small, random discrepancies in one EOF will propagate into bigger differences at higher order because each additional EOF is orthogonal to all others. "*

Groot et al., 2026 (JEMS) shows that the variability is actually highly similar. However, we have rephrased this to "has large similarity", because this work indeed indicates that 1 EOF looks almost identical, which represents the dominant contribution to the free-tropospheric variability (and almost half of the variability in our dataset).

39. *"l. 605f again, I don't know how you think that the EOFs make any statements about any kind of mean when they are derived exclusively from the covariances. "*

We have revised the text, because this discussion point is about the advanced multi-variate links between tropical CAPE change and model tendencies. Buschow (2024, QJRMS) suggested that the link might be simple: CAPE and convective precipitation could be linked. Some statistical support is shown for this hypothesis. We indicate that initialised IFS (from some analysis or parent simulation) likely has more complicated links between CAPE change and model tendencies, broadly independent of precipitation. The text has been revised for clarity (revised lines 654-658):
Our analysis suggests that Buschow's (2024, QJRMS) broad latitudinal band has probably missed associated multi-variate adjustments comprehensively, which requires multiple EOFs to represent $\delta CAPE$, also in IFS. In our experiment $\delta CAPE$ is mostly associated with mixed-layer adjustment. After starting SCMs with an unbalanced ICON benchmark state, we find mostly tendency variability that decouples precipitation and $\delta CAPE$ within our EOFs. In fact, at large scales, the mixed-layer drying tendency appears to explain our $\delta CAPE$ variability very well.
The following lines have been removed. The text is intended to be simpler and still correct and complete.

40. *"l. 608 "orthogonality between CAPE and precip" please clarify which part of your paper you refer to here. The EOF analysis included only CAPE tendency, not CAPE itself. Did Buschow (2024) discuss CAPE or CAPE tendency?"*

CAPE tendency was discussed. We removed this phrasing to avoid possible confusion when adjusting for the previous comment of this reviewer (number 39) and aimed to clarify the contrast

between Buschow (2024) and our analysis by addressing it.
We thank the reviewers for the long list of suggestions, their persistent efforts and specific questions (and the editor for reading and digesting the comprehensive material). We hope to have clarified each point individually and thoroughly with the above replies.

The references are all included in the reference list of the revised study.