Reviewer #3

The study seeks to assess improvements to the prediction of tropical shallow cumulus regimes by modifying CLUBB to allow for counter-gradient momentum fluxes. Overall, it is important work and I found the paper well written and straightforward to follow. I have a few comments that require revisions before the article should be published, but I consider them minor. I have attached specific comments in a PDF.

Thank you for your positive comments and subsequent feedback. Please see responses below.

Lines 30-33: "Changes in low cloud fractions… The Hadley cell" – it's unclear how the Hadley cell links to the opening/topic sentence of this paragraph. Suggest adding language to more cleanly transition between the two sentences.

Agree this transition was awkward. This has been reworded.

Lines 149-152: Was the CMIP6 version using 58L CAM, with the refined resolution in the BL, and with the SE grid? My impression is that this might be different from the original CESM2 release version.

Yes, you are correct, we use a version that differs slightly from the official out-of-the-box CAM6 release -- mainly that we use the spectral element dynamical core instead of the finite-volume and 58 vertical levels instead of 32. To clarify this for the readers we have changed this to read:

"The version of CAM studied here is CAM version 6 (Bogenschutz et al., 2018; Gettelman et al., 2019). This corresponds to the configuration of CAM in the CESM version 2 release (Danabasoglu et al., 2020) that was used to generate the simulation submitted to the Coupled Model Intercomparison Project version 6 (CMIP6), with two differences. First, we use the spectral element (SE) dynamical core (Lauritzen et al., 2018) on an unstructured cubed-sphere grid with nominal 1∘ (111km, also referred to as CAM-SE's ne30np4 grid) horizontal grid spacing. This is in lieu of the CAM6 default finite-volume dynamical core. Second, we use 58 vertical levels with finer grid spacing in the atmospheric boundary layer compared to CAM6's default 32 layers."

Lines 167-168: "C6 and C7 are also tunable constants, although they are left as 4 and 0.5, respectively, for all simulations here." – are these the default values of C6 and C7 that CLUBB uses out of the box?

Yes, we have modified the text to better clarify this, but to answer here, we preserve the default CLUBB settings in the CAM6 release in the "O" runs.

Lines 236-238: "At each of these 10-meter levels, state variable values meant to represent model output are calculated…" – the phrasing here is a bit confusing. The state variables are the model output, no? Is the point ultimately that model output is interpolated as a linear vertical distance-weighted average for every 10-meter observation?

To clarify this we have replaced "… [at] each of these 10-meter levels, state variables meant to represent model output are calculated by taking the linear vertical distance-weighted average of those values reported at the nearest two model levels." with "… state variables from model output are linearly interpolated to each of these 10-meter levels by taking the linear vertical distance-weighted average of those values reported at the nearest two model levels."

Line 284: "This along with observations in our study being qualitatively similar to the LES-derived profiles in L19…" – Is this implying that the limited observations of u'w' are in line with the LES profiles of L19? I'm confused by the use of "observations" here, which seems contrary to what was stated in the paragraph before.

"Observations" in line 284 refers to observations of u and v, not <u'w'> like the bulk of the paragraph. We acknowledge this was unclear and have updated the sentence to read: "*This along with observations of u and v in our study being qualitatively similar to those in the LES-derived profiles in L19…"*

Lines 290-291: "In x101, v'w' is also about half as negative at altitudes between 300 m and 2 km." Would be helpful to note that this refers to Figure 1d, not 1c.

We have added a figure reference to this sentence to refer to Fig. 1d.

Lines 302-310: I'm not sure what the discussion of the Ekman spiral in the atmosphere lends to this study in particular. Perhaps draw a clearer link or consider removing most of this?

Removed.

Figure 2: It seems that panels (a) and (b) are just repetitions of Figure 1 (a) and (c); is there a way to combine them then to reduce redundancy? Would also be good to name in the caption which panels refer to which part (i.e., "Vertical profiles of means (a-c),…") even though the axis labels are fairly clear.

The reviewer is correct that those two panels are reproduced. We cannot come up with a clean way of eliminating them from Fig. 2, however, we add text that notes they are identical to part of Fig. 1. We also like the idea of adding parentheticals regarding the panel labels and have amended the caption to include the suggestion.

Lines 314-315: "It can be seen that although x101 has a stronger jet maximum than x001, it has a reduced maximum easterly bias when compared to x001 since its jet placement matches observations better." – this feels redundant as well, since the smaller bias maximum was noted when discussing Fig 1. Combining Figs 1 & 2 might make this a bit easier to discuss with less repetition.

The text has been amended to reduce the redundancy. We also specifically call out individual panels in Fig. 2 when discussing the bias and RMSE results to focus the reader's attention as they make their way through the paragraph.

Line 317: “near the jet maximum” – is this near the observed jet max, or the model simulated?

After reviewing the figure, “*near the jet maximum*” has been replaced by “*in the region immediately above the modeled jet maximum (roughly 1 to 2 km altitude).”*

Lines 317-318: “The remainder of the RMSE profiles are quite similar…” – they’re nearly identical for v, but for u it looks like the similarities are fairly different throughout the vertical; maybe a more nuanced statement is warranted?

True, *“[the] remainder of the RMSE profiles are quite similar*,” has been changed to *“[both] the RMSE profile for v and the RMSE profile for u far from the modeled/observed jet maxima are quite similar.”*

Figure 3: “The vertical axis is a rough estimate of the pressure level of the model output” – could you be more specific? Is this the hybrid coordinate pressure?

To better specify, we have added a note *“... levels here are taken from a column at a single time, making the pressure levels estimates, since the hybrid pressure coordinates change depending on elevation and surface pressure. In this situation, this is a reasonable estimate since all balloons were launched from near sea level and almost all drifted over the open ocean in fair weather conditions*.”

Lines 335-336: “Most points with negative Keff in x101 are above this threshold…” – how far above the threshold do these points typically lie? Is there a large spread in the value, and values are often much larger than the threshold, or are values often close to the value (and perhaps thus the findings are sensitive to the choice of cutoff)?

We have included a histogram of K_{eff} values in Fig. 10. The number of points lying below this threshold is 0.1-0.2%. When this filter is included, the mean (median) wind shear for negative K_{eff} points is 0.52 (0.43) m/s per km and 2.2 (1.6) m/s per km for PM-O and PM-X respectively, indicating the majority of the negative K_{eff} points are generated in the presence of wind shears at least double this threshold.

Lines 347-348: “Confidence is added to this hypothesis by…” – I see the discussion of LES results with different forcing (i.e., Helfer et al.), but does this refer to other studies that use the EUREC4A/ATOMIC forcing? Would be good to discuss/cite those if so.

We have added some additional references to recent work using LES with ATOMIC/EUREC4A data, albeit primarily focused on shallow convective cloud structures.

“We also refer interested readers to Narenpitak et al. (2021), Dauhut et al. (2023), and Schulz and Stevens (2023), all of which performed LES simulations using a variety of configurations to investigate the distributions and organization of shallow convective clouds during the EUREC4A/ATOMIC study period.”

Figure 4 and related discussion: Are these differences in theta and Q profiles statistically significant?

Yes, even though the differences are quite small, they are systematic across a vast number of soundings. To test this, we performed a paired t-test at each altitude in the sounding across an ordered list of the soundings in the observational comparison. The majority (>50%) of altitudes were different at alpha = 0.05 level (some altitudes did not register as statistically significant such as θ just below 1 km, where the profiles lie nearly on top of one another.

To highlight this in the text we add: "*These differences in thermodynamic profiles are not as large as the differences in the momentum profiles but do exist. In fact, these differences are still significant at most altitudes when performing a paired Student's t-test across the model profiles included in Fig. 4 (92% (72%) of altitude bins in the θ (Q) profiles significantly differ between ED-O and PM-O at the α = 0.05 level).*"

Line 382: A better qualitative match, yes; is this not a better quantitative match as well?

Yes, the experimental length scale runs match observations of v better at most altitudes both in terms of absolute error and in structure, the word "qualitative" is removed from this sentence.

Lines 382-385: Suggest adding an in-text reference to panels of Fig 6 as they're discussed. In terms of reducing theta/Q biases in the x200 runs, isn't this to be expected when any run is tuned to better match the ERA5 results? It seems that potential bias reductions in these runs could be driven more by the tuning than by the formulation of L/prognostic momentum.

Figure panel references have been added in this section. Although the model configurations were tuned to match ERA5 state variables, we only attempted to match *u* and *v* globally. It was not known ahead of time if small variations in these parameters would noticeably affect model predictions both for thermodynamic quantities and quantities in the study region.

Figure 7: It looks like the experimental L formulation has a substantial impact on u biases in the lowest 1 km. In 001 and 101, negative biases extend to the surface, but those seem to be removed with the L cases. Is there a reason for that? Is the L formulation most sensitive closest to the surface?

The bias reduction in u in the experimental L cases can be seen in Fig. 5 as a rightward shift of the mean wind profile in the experimental L runs relative to the eddy diffusivity (ED-O) and original length scale prognostic runs (PM-O). This bias reduction is most notable relative to PM-O below 1.2 km where PM-O overestimates the strength of the jet maximum, and relative to ED-O above 1.2 km where ED-O diffuses momentum too high. The bias reduction appears to be strongest in the lowest 1 km in Fig. 7 likely because this region has lower bias to begin with and thus the same linear reduction is a larger percent reduction, and because the color bar begins to change from red to blue making the reduction more noticeable.

Line 404: "Tropospheric" should probably be "troposphere".

Correct, this has been updated.

Line 410: "between 200 and 2 km" – should read 200 m and 2 km.

Thank you for catching this. Corrected.

Figure 8: Missing legend

Fixed.

Lines 424-425: "do demonstrate a likely connection between the prediction of upgradient fluxes and modifications to various terms in the vertical momentum flux budget" – Could you elaborate on this a bit more for clarity? How does this tie into the vertical momentum flux budget terms? It seems that this is just the prediction of upgradient occurrence in the figure.

We agree this was unclear. It has been reworded to read: "*We emphasize that these more frequent predictions of upgradient fluxes are not necessarily more accurate, however, they do demonstrate a likely connection between the prediction of countergradient fluxes and modifications to the turbulent dissipation in CLUBB. That is, in the `PM' simulations, changes to the turbulent length scale aimed at improving the shape of the near-surface u and v profiles can further enhance the generation of upgradient momentum fluxes.*"

Fig 10: Would be helpful to have additional percentages labeled, not just 100% and "same" (and perhaps same should be written as 0%?). Overall the colorbar combined with the actual bias values in the boxes is a little confusing. It would seem for example that the bias in x101 for Mixing Ratio should be not quite the darkest red (it's not a doubling of the bias), but It's the same color as x201, which is more than a doubling of the bias…

We have relabeled "same" as "no change" in response to this comment – we agree that fits better.

We have also added additional labels to the bar rather than just the minimum, maximum, and central points as before.

Reported values here are rounded to two decimal places, so moving from ED-O to PM-O (from -0.02 to -0.03) is not actually a 50% increase in bias, but indeed a greater than 100% increase in bias (from -0.016 to -0.033). We agree that the coloring is not very informative for the mixing ratio row, but this is described in the text. We have considered a few other options such as standardizing by the mean value, although this artificially scales quantities based on their mean state (e.g., temperature). To preserve the color scale while maintaining consistency across all variables, we choose to leave the color bar as is and further emphasize this fact in the text.

Lines 445-446: It's worth noting that although "the greatest improvements are seen in u and U_h, there's a stronger degradation in Q when you add in the experimental L calculation. Would be a more balanced description of the results, at least; elaboration would be great.

We have added the caveat "*Some bias degradation is seen in these means for v and Q, but these results are not very meaningful as the mean biases for both these variables are small to begin with and the absolute changes in biases between model configurations are small as well*."

Fig 11: Please add additional colorbar markers, as for Fig 10.

Done.

Lines 489-490: "One of the most notable…" – suggest adding a parenthentical reference to guide the reader exactly where to see this. So maybe at the end, add "(solid brown line in Fig. 12)"? Would help in additional sentences of this paragraph as well.

Thank you for the suggestion. Parentheticals describing the line styles in the figure have been added to this paragraph.

Lines 495-496: "…could be leading to changes in atmospheric stability…" – any evidence that could be added to support this?

We do not have direct evidence, although the thermodynamic profiles in Fig. 6 show changes in $d\theta/dz$ in the different runs (static stability). More constrained models (e.g., single-column or nudged simulations) may help shed light on exactly the mechanics at play, although that is beyond the scope of this study.

We have added a pointer in the text to the θ profiles in this section. We also add "*We admit this is speculative, however, and experiments with more constrained model configurations (e.g., single column, nudged runs) would be helpful in providing deeper insight*."

Lines 528-536: "This study is a targeted regional investigation and as such, the improvements seen here cannot necessarily be generalized to the global climate system without further exploration…" – This is a really important caveat, and I appreciate the discussion surrounding it. The question arises then – why not use these simulations to evaluate global performance? You have the full global output, so could this dataset be a tool for exploring additional regions/field campaigns, and more generally for looking at global biases? It may be beyond the scope of this particular study, but is it something that's targeted for future work or are the runs not suitable for that analysis

The reviewer is correct in that there is nothing preventing the global results from being evaluated. Some global data is in fact included in the data DOI attached to this manuscript for other researchers who'd like to explore. It's worth noting that some of the specific turbulence quantities used to evaluate the momentum flux budgets are only output over the EUREC4A/ATOMIC study region to reduce the data output burden while the model was being run. However, we do note that we hesitate to undertake a detailed global evaluation given the fact that evaluation and tuning are specifically targeted on the EUREC4A/ATOMIC region and our experience indicates that attempting to optimize tuning parameters for a particular geographic region will almost certainly result in at least some degradation elsewhere in the global simulation (also see Hourdin et al., 2017, BAMS). That said, ultimately the goal of this work is to have some of the updates described in this manuscript implemented in the global version of CAM used for CMIP-class simulations, so a more detailed understanding of the turbulence budgets in multiple atmospheric environments is an (exciting) target for ongoing and future work.