Responses to Reviewer 1

Dear Reviewer, we appreciate your time and effort in acknowledging and thoroughly reviewing our manuscript. We are truly grateful for your constructive comments and insightful suggestions, which encourage and help us to improve the manuscript. We have revised the manuscript carefully based on your comments.

In the responses below, your comments are provided in black text and our responses are provided in blue text.

This work studies aerosol-cloud interactions in the Eastern North Atlantic region, comparing satellite retrievals to a GCM model output, for a large set of days, to elucidate the relationship between micro- and macrophysical cloud properties: cloud droplet number concentration, liquid water path, and boundary-layer extinction coefficient. They analyze the relationships between these variables seasonally, comparing simulations to satellite data, and the main novelty of the study is that they also analyze the behavior for 4 meteorological regimes that are found using clustering techniques on ERA5 reanalysis data. This regime clustering gives new insights by separating natural covariability and clarifying one of the relationships. The paper is well written, and the discussion is very detailed and provides a full understanding of the studied system and its physical processes. I mostly have minor comments regarding some methods, and about how to better summarize and provide ideas to modelers based on their discussion.

We sincerely appreciate your thoughtful and constructive feedback. All comments have been carefully considered, and the manuscript has been revised accordingly.

Minor comments

• I suggest highlighting the novelties of the paper in the abstract, introduction, and summary. In particular, I am not sure if the novelty is only the analysis based on regimes, or if the seasonal analysis is also novel? Or is this particular model and satellite product comparison new?

The primary novelty is the regime-based evaluation framework that places E3SMv2 and satellite observations side-by-side within the same synoptic regimes, allowing like-for-like attribution of ACI behavior. A second novelty is our use of a new vertically resolved aerosol-extinction product to diagnose free-tropospheric versus boundary-layer influences on the LWP-N_d relationship. The seasonal analysis is included chiefly to reassure and reconfirm prior findings and to show that the regime-based conclusions are robust across seasons.

We have revised the abstract, introduction, and summary to state these contributions explicitly.

• The discussion in every Section is very thorough, but many hypotheses point to modeling biases or ideas for model improvements, which are not the main scientific contribution of this work. It would be nice to assess if these hypotheses are true by confirming some diagnostics on the resulting model parameters. Another thing that could be done is to order these recommendations and try to assess which model improvements are more likely or feasible.

We appreciate the suggestion to verify our hypotheses with targeted parameter diagnostics and to prioritize feasible improvements. However, our study's main contribution is to provide a novel regime-based evaluation which helps narrow down the conditions where the model uncertainties in ACI are the largest and further identify the processes associated with those specific regimes. Analyses of comprehensive parameter perturbations require a set of paired simulations that are beyond the scope of this work. Instead, we carefully grounded our attributions to published model sensitivity and resolution studies. Previous sensitivity studies indicate that changes in low clouds in E3SMv2 are primarily controlled by CLUBB, followed by MG2 tunings (Zhang et al., 2023). Independent analysis of the MG2 scheme show biases in warm-rain processes and that it realizes the negative LWP pathway too rapidly and strongly (Zhou et al., 2025), implicating the turbulence-microphysics coupling as a persistent bias source that aligns with our regime-specific over-depletion of LWP at high N_d. Moreover, model resolution studies found that vertically resolved physics and concurrent horizontal and vertical refinement improve the representation of entrainment mixing processes and reduce stubborn stratocumulus biases (Lee et al., 2022; Bogenschutz et al., 2023). Hence, we cautiously attribute the potential E3SMv2 discrepancies

versus satellite results to those simulated processes in the model, while acknowledging that they can also be the combined effects of multiple feedbacks and interplay among the model schemes.

In terms of the feasibility of potential model improvements, we think that a feasible approach would be the fine-tuning of the microphysical parameterization, ideally constrained by high-resolution observational data from field campaigns such as ARM. This may reduce the persistent uncertainties in simulating aerosol-cloud interactions, particularly under the dynamic meteorological transitions typical of the ENA region. Furthermore, emulation from high-resolution modeling (e.g., LES) of cloud and rain microphysics processes can be used to replace the bulk microphysics scheme, which can contribute to better performance with manageable cost as shown in previous studies such as Gettelman et al. (2021). Increasing spatial resolution is also feasible in a regionally refined mesh, and increasing vertical resolution might follow, but both would noticeably increase computational cost, so trade-offs should be considered with caution. Lastly, the development of new schemes that bridge the gap between shallow and deep cloud regimes remains particularly challenging, as current large-scale model schemes still treat them separately.

We have added the above discussions in the revised Section 5.

Line by line comments

- L17 Clustering was performed on satellite or simulation data? Or both?

The clustering is applied on the ERA5 reanalysis, then the satellite and model data are aggregated based on the clustered regimes.

- L18 Maybe explain the 4 regimes before they start appearing

We have revised this statement to ‘We then partition ENA meteorology into four synoptic regimes (Pre-Trough, Post-Trough, Ridge, Trough) via a deep-learning clustering of ERA5 reanalysis fields’.

• L161 Are there comparisons for other cloud types?

Yes, Gryspeerdt et al. (2022) explicitly compares satellite-retrieved N_d with in-situ aircraft data across multiple cloud regimes. They find high fidelity in marine stratocumulus and lower correlations in more challenging convective situations. And this is precisely why, in our study, we prioritize low-level liquid clouds, where the satellite retrieval is best-validated and most defensible for model-satellite ACI evaluation.

We have revised the statement to '…previous studies have shown that the N_d compares well with measurements from 11 aircraft campaigns, demonstrating a decent correlation when sampling the marine stratocumulus clouds, with r^2 values of 0.5~0.8 (Gryspeerdt et al., 2022). Therefore, to minimize known retrieval uncertainties, we focus on low-level liquid clouds where satellite N_d shows the strongest aircraft agreement and typical normalized root mean squared deviation of ~30-50 % (Gryspeerdt et al., 2022).'

• L195 What is the value of that coarse vertical resolution?

Our E3SMv2 simulation employs the standard ~72-layer atmosphere, giving a near-surface vertical resolution of roughly 50–100 m, gradually coarsening to ~200–300 m per layer near cloud top through the free troposphere.

For a rough estimate, we have revised the statement to 'Given the coarse vertical resolution of E3SM near the cloud top (~200-300 m), …'

• L213 Time formatting: Should it be 1 p.m or 13:00?

We have changed the occasion to '13:00 LT (1 p.m. local time)' for more consistent formatting.

• L214 Was the date also a variable?

We did not include calendar date (or month-of-year) as an input feature; the model ingests only Z500, SLP, and 10-m winds, using temporal ordering (not absolute time) for the LSTM. Since

adding date would impose a seasonal prior that can bias the clustering toward calendar timing rather than physical flow patterns.

• L226 So the DEC was used after optimizing the k-means clustering? Or was it also tested for different k values?

Yes, the DEC was used after optimizing the k-means clustering. We first determined the optimal number of clusters *k*, and then ran DEC with that fixed *k* (4 in this study). DEC was initialized with the k-means centroids and then optimized the KL-divergence (KL) clustering loss between the encoder's soft assignments and a sharpened target distribution, with periodic centroid updates, and the *k* remained the k-means-optimized value throughout.

We have clarified the methodology as follows:

'…To further refine the clustering assignments, we then ran DEC with that fixed cluster number of four, as determined with K-means optimization. DEC was initialized by the K-means centroids and optimized the KL-divergence clustering loss (between soft assignments and a sharpened target distribution) with periodic centroid updates…'

• L246 "followed by fall", "lowest during winter"?

Thanks for the correction. We have revised it to 'followed by fall (SON, 73.36 cm^{-3}), and the lowest during winter (DJF, 60.37 cm^{-3})'

• L278 Is this index computed from the data? Is it a fit with confidence interval?

Yes. The adjustment index $\mathcal{L}_0$ is computed directly from the data for satellites and for E3SMv2 separately.

The reported "±" values are the standard error (SE) of the slope of ordinary least squares fit in log–log space (scipy.stats.linregress). A 95% confidence interval can be reported as $\mathcal{L}_0 \pm 1.96\times$SE. Here our main text currently shows the slope ± SE.

We have clarified the methodology as follows:

'We compute $\mathcal{L}_0$ as the slope of an ordinary least squares fit in log-log space between N_d and LWP. Hence, the $\mathcal{L}_0$ derived from satellite observations and model simulations is -0.192 ± 0.006 and -0.375 ± 0.005, respectively. The '±' values reported are the standard errors of the slope (SE) from that fit (equivalently, 95% confidence level CI = slope ± 1.96*SE, under standard linear-regression assumptions).'

• L355-359 This sentence is a bit confusing

We have revised the discussion for better clarity:

'Aircraft in situ measurements near cloud base provide the most physically robust ACI assessment (Gupta et al., 2021; Zheng et al., 2024). However, it is challenging to do that with satellite data and model outputs, because satellite remote sensing like CALIOP cannot reliably determine cloud-base height, and the model's coarse vertical resolution makes it difficult to collocate the model cloud-base with CALIOP layers. Hence, those factors necessitate the use of the mean aerosol properties within the below-cloud-top MBL in the present study'

• Fig. 5: Composites mean that these are based on the mean values of each cluster? Or are these the centroids?

& • L401 Details were already given in the previous Section

For each regime we average the ERA5 fields over all dates classified into that regime to form the maps shown (Z500 with winds, SLP with winds, and LTS). Thus each panel represents the mean state of all members in that cluster.

Hence, we have revised description of Figure 5 for better clarity below:

'As detailed earlier, the CNN–LSTM–DEC clustering of 3,286 daily ERA5 states results in the identification of four distinct synoptic - scale regimes (Figure 5). Namely, Pre-Trough (regime 1), Post-Trough (regime 2), Ridge (regime 3) and Trough (regime 4). For each regime, composites

were computed as the arithmetic mean of the corresponding ERA5 fields across all time steps assigned to that regime.'

• L407 Is there a reason why the regime order does not follow the expected trough-ridge transition?

We decided to pair Ridge with Trough in figures and discussion to create a clear side-by-side contrast, and Pre-Trough with Post-Trough because they have comparable sample sizes and bracket the Trough disturbance. And the regime labels are permutation-invariant outputs of the unsupervised clustering. We order them for readability rather than chronology, so numbering should not be interpreted as a trough–ridge time sequence.

• L426 I think it is important to report the number of events and percentage for each regime in the main manuscript, for statistical significance. Now that I see the supplementary information, maybe it is worth cautioning the readers that regime 4 had the lowest amount of information

Thanks for the suggestion, and Reviewer 2 also raised similar comments.

Hence, we have moved Table 1 to the main text. And added the following statement: 'Note that among the four regimes, Regime 4 is the least frequent (3.4%) and is largely confined to the colder seasons (winter and spring), confirming the findings from previous studies'

• L442 "are listed"

Thanks, correction has been made in the revised manuscript.

• Fig. 9: The median sigma values were selected for each regime or for the entire dataset?

We used a single median threshold for σ_{MBL} computed from the pooled satellite and E3SM dataset rather than regime- or dataset-specific medians. This ensures an identical conditioning for both data sources and all regimes, avoiding different bin edges that could confound interpretation.

We have clarified that as follows:

'In order to further illustrate the impact of aerosols on the behavior of the LWP–N_d relationship, both satellite and E3SMv2 data are grouped into lower and higher half σ_{MBL} categories, defined by the pooled median of the combined satellite and E3SM samples (0.594), and this single threshold is applied to all regimes and both datasets to ensures an identical conditioning (Fig. 9).'

• L720 I suggest mentioning the four regimes

The four regimes are now mentioned in the beginning of this paragraph.