Authors would like to thank *Anonymous Referee #1* for the review.

The authors present a detailed study of the ability of a collection of models to reproduce the in-situ and remotely sensed properties of the biomass burning plume obtained as part of the ORACLES 2016 campaign over the southeast Atlantic. They show that the campaign sampled a relatively representative portion of the plume in space and time. They find that the models tend to underestimate the height of both the base and the top of the plume against these observations, and that most models underestimate the mass extinction efficiency within the plume.

While the paper is well written and comprehensive in its analysis I feel the results need to be put into a broader context and include deeper interpretation for it to fall within the scope of ACP. For example, it isn't clear what the implications of the highlighted biases are in the, fairly arbitrary, selection of models chosen.

The summary is missing an assessment of the impact of the underprediction of the modelled plume heights on e.g. the local aerosol forcing through direct and semi-direct effects. This could be linked to recent work by Gordon et al. 2018 more closely, especially as the same model was used.

As discussed throughout, the modeled extinction and SSA values are diverse in comparison to the observations. The direct and semi-direct effects also depend upon the properties of the underlying cloud field, which are beyond the focus of the current manuscript and treated within an upcoming companion paper by Doherty et al. This is now stated within the Summary. That said, we have included a section within the Discussion (Section 7.3) that discusses how the documented biases might affect the model estimates for the aerosol radiative effects, reproduced below in a contrasting font color:

"7.3. Impact of model biases upon calculated aerosol radiative effects

The ultimate goal of this study is to provide groundwork towards improving the physically-based depiction of the modeled aerosol radiative effects (direct, indirect and semi-direct) for this

climatically-important region. Zuidema et al. (2016) indicate a wide range of modeled direct aerosol radiative effect (DARE) values for 16 global models. Similar to this study, no standardization was imposed upon the model simulations. Of these, the GEOS-Chem model is also represented within this intercomparison, with the caveat that some model specifications may have evolved in ways we are not aware of. The CAM5 model is also incorporated within the WRF-CAM5 regional simulation of the current study, using the same MAM3 aerosol microphysics. GEOS-Chem reports a small but positive August-September DARE (+0.06 W m-2) and the global CAM5.1 model reports the most warming (+1.62 W m-2) of the 16 models shown in Zuidema et al. (2016).

The current study does not assess the model cloud representations other than WRF-CAM5 cloud top height, upon which all the aerosol radiative effects also depend. Most models, including GOES-Chem, WRF-CAM5 and ALADIN-Climate, share the bias of generally underestimated BC mass within the 3-6 km layer offshore, and overestimates closer to the coast. Although speculative, the weakly positive DARE within GEOS-Chem is consistent with a GEOS-Chem overestimate in ACAOD that is compensated by its SSA overestimate, all else equal. The EAM-E3SM model biases are similar, and suggest similarly compensatory behavior will impact the model DARE estimates. The more robust performance of WRF-CAM5 within this intercomparison, if that can be extrapolated to the global CAM5, would imply support for the more strongly positive global CAM5 DARE estimate relative to the other models within Zuidema et al. (2016).

ALADIN-Climate is a regional model reporting a more positive top-of-atmosphere DARE of approximately 6 Wm-2 over the ORACLES domain for September, 2016 (Mallet et al., 2019) than any of the global models. Reasons for this are beyond the scope of this study, but the ALADIN-Climate underestimate of ACAOD combined with a slight SSA overestimate suggest that the ALADIN-Climate DARE is likely still underestimated. Mallet et al. (2020) investigates the model sensitivity to smoke SSA, and finds a variation of 2.3 Wm-2 that can be attributed solely to SSA variability, for July-September DARE. The UM uses a two-moment aerosol microphysics scheme that is updated from the one applied within the HadGEM2 model of de Graaf et al. (2014), and no

UM DARE estimates are yet available. The EAM-E3SM incorporates a sophisticated new MAM4 aerosol scheme that explicitly includes the condensation of freshly-emitted gases upon black carbon. The EAM-E3SM results within this study use a long-term monthly-mean emission database, and future work will examine model DARE values specific to September, 2016. An upcoming companion paper will include all of the variables needed to calculate DARE, allowing for a more quantitative evolution of the model bias propagation."

We have also edited the manuscript to further emphasize what we consider the strengths of the study: a focus on the spatial distribution of a wider range of aerosol composition and optical properties than has previously been done. We have, however, added an additional figure indicating that the modeled heights of the low clouds typically exceed those observed – indicating that it is easy for the models to overentrain biomass-burning aerosol into the boundary layer."

We do note that the model used in Gordon et al., 2018 is not analysed here, as we only use the global model that provides the boundary conditions to that study.

The paper would also greatly benefit from a clearer focus to help guide the results section which becomes quite hard to follow otherwise.

We have made a number of significant edits to provide a clearer focus. This includes a clearer emphasis on comparisons within the free troposphere, for which we can say more than those within the boundary layer. Please see the revised document for the changes.

In particular, the link between the biases in the aerosol microphysical and optical properties isn't elucidated until the discussion. Even then I feel the discussion isn't placed in sufficient context: There is a large amount of diversity in model estimates of the absorptivity of the plume in the literature and the comparisons here could go a long way to unpicking this.

We have included more discussion of the links between the biases in the aerosol microphysical and optical properties in the Discussion section. A clear result is that most models overestimate the amount of organic aerosol mass relative to that of black carbon. This will have implications for the single scattering albedo and our proxy for the mass extinction efficiency. Other model biases are more diverse, with different model processes likely responsible in each model. Ultimately the modelling centers responsible for the individual models will need to uncover these processes. The intent of this contribution is to support that activity.

Other, more minor comments and suggestions are provided in the attached PDF.

ALADIN-Climate, for which an extinction threshold of 17 Mm-1 is used - why not 15 like the observations?

It is because 15 Mm-1 is for the observations of dried particle whereas the ALADIN-Climate threshold is defined for ambient extinction. The manuscript now says “an extinction threshold of 17 Mm^{-1} at ambient relative humidity, which approximately corresponds to 15 Mm^{-1} at low RH, is used”.

A MBL is not defined for the HSRL data?

HSRL-2 gives cloud top height, with which one could define MBL. Our paper does not identify MBL this way because it excludes the locations without clouds and because extinction measurements are not available below optically thick clouds. We have, however, added a new figure (Fig. 16) that compares the HSRL-2 cloud top height (CTH) with WRF-CAM5 CTH as well as the boundary layer height from each model.

How might the different re-analysis products used to drive the large scale dynamics in the models contribute to these differences [between the observed and modeled variability in smoke heights for southernmost boxes]?

We refer to differences in the driving meteorology as one of several potential causes for the differences in Sect. 7. Beyond that we can say little about the difference among reanalysis products.

I don't feel showing the ambient diameter for the UM adds anything to this discussion and just makes interpretation harder: The modeled diameters are 20 % greater in the ambient RH. [Commented in the main text.] The observations are dry diameters so only the UM dry results should be shown [Commented on Fig. 8.]

We have removed the ambient values from Fig. 8 and modified the text accordingly.

How does this [WRF-CAM5's a prescribed volumetric geometric mean diameter of 375 nm] compare to the emission size used in the UM?

The manuscript now says "[…] compared to the UM's 228 nm. Note the volume (arithmetic) mean diameter is smaller than the volume geometric mean diameter."

It doesn't seem fair to include ambient extinction against dry observations. I think you should just show the only model to give you dry (or not at all). [Commented in the main text.] Again, only the model values at the correct humidity should be compared for this and the following plots, they're impossible to interpret otherwise [Commented in Fig. 10.]

For the free troposphere the observed impact of hygroscopicity is very small. As Section 6.2 says, the ambient-RH/dry ratio of light scattering is estimated to be less than 1.2 for the 90 % of the time when the dry scattering exceeds 1 Mm^{-1}, according to concurrent, once-per-second measurements with two nephelometers with instrument RH set respectively to high (~80 %) and low (~20 %). We therefore find merit in the model-observation comparisons without the adjustments for humidity differences. Some models seem to have greater hygroscopic effects internally, however.

In the marine boundary layer the hygroscopic effects are significant. We discuss it referring to the *in situ* hygroscopicity measurements in Section 6.2. In addition, we have inserted two papers that highlight overestimates in the GEOS-5 sea salt emissions.

What refractive indices do the models use? Could this explain some of these discrepancies [in MEE, the mass extinction efficiency]?

The diversity in the model biases of the extinction to OA+BC mass ratio suggests different processes may be responsible for the biases in each model. While their attribution is beyond the scope of this study, we hope that documenting the biases in both MEE simultaneously with those in the underlying aerosol properties will aid future process attribution studies leading to improved parameterizations. We do not know the refractive indices of the individual aerosol components and how these are combined within the individual models.

This [Table 3] is very hard to read and might be better as a graph.

The descriptions of the inter-comparison results and discussion now center on the figures. Table 3 and Table 4 have been brought to the supplementary material.

An explicit formula for volumetric arithmetic mean diameter

The manuscript now includes the formula for the volumetric arithmetic mean diameter of the accumulation mode. This is $(V/\pi*6/N)^{1/3}$, the cube root of the volume-to-number ratio (V/N, where V and N are integrals of the volume and number over the UHSAS diameters for each size distribution) after the volume is divided by $\pi/6$".

Page 10, line 20, form should read for.

The original sentence mentioning future intercomparisons has been dropped to give clearer focus.

This doesn't quite make sense, consider re-wording: An initial evaluation of the free-tropospheric aerosol layer top and bottom altitudes 6 prepares for the comparisons carried out for the comparison layers.

Re-worded to "Here we provide an evaluation of the free-tropospheric aerosol layer top and bottom altitudes, in preparation for the comparisons of the vertically resolved values."

Insert the before smoke layer top, at around before 5-6 km.

Inserted.

It would be nice to have this in Km too, for consistency: 1740 ± 290

The text now says "The zonal gradient in observed plume top and bottom heights along 8° S is small (Fig. 5b), with mean altitudes +/- standard deviations between 3° W and 13° E of 5.25 km +/- 180 m and 1.74km +/-290 m respectively. ." Altitudes are expressed in km, and their differences and errors in m.

And in the vertical? : the location and time

"(in the vertical and horizontal)" has been inserted after "space" in Section 4. And "location" has been replaced with "space" in Section 6.

Perhaps don't include this plot [MBL SSA comparison, which is subject to poor statistics] then - it seems a bit unfair on the models.

We elected to keep the figure to be consistent with the other figures. The figure caption now emphasizes the lack of adjustment for the humidity effects. "Note that the modeled SSA refers to the ambient humidity whereas the observations are for dried particles." As part of the manuscript edits, we have more strongly emphasized the comparison within the free troposphere, where it is more robust.

I would suggest only including statistics [in Fig. 5 and probably in all other box-whisker plots] which include a minimum number of samples, at least 10, to ensure the statistics are at all representitive

While this suggestion seems reasonable to ensure the representativeness for each property, it would complicate the interpretation of the link between multiple variables and likely aggravate the regional representativeness. The number of observations for a given box and for a given altitude differs between properties. For example, for the northernmost box of the meridional corridor at 3-6 km, 4-5 mass measurements are available compared to 15 P3-borne measurements of *in situ* optical properties. A threshold of 10 samples would exclude the masses but keep the optical properties. This would make the interpretation of the link between them (e.g., MEE) more complicated than it already is. To minimize this impact, the threshold for optical properties would

have to be set higher. There is no easy way to determine exactly how high it should be, as the sampling rate varies from box to box, from altitude range to another. And, even if one manages to settle on a threshold for every property, the result would likely exclude many combinations of box and altitude range. Thus, the pursuit of better statistics within each box and altitude range would result in poorer representativeness across the study region and altitude ranges.

We do recognize the issue, and by including the number of samples for both the observations and models on each comparison figure, provide the information needed for individual readers to discriminate. We also focus on the more robust comparisons within the text.

The y-axis labels [in Fig. 11 and 12] should be shortened or split on to more lines to avoid them clashing.

Shortened.

These panels [in Fig. 15, and others] are missing (a, b, c…) labels.

Inserted.

Authors would like to thank *Anonymous Referee #2* for the review.

This paper presents a statistical comparison of aircraft observations of smoke aerosols along repeated sampling tracks from the 2016 deployment of the ORACLES campaign against a variety of model simulated aerosols for grid cells along the same sampling tracks. Few field campaigns provide sufficient sampling to allow for such a comparison and the authors go to some lengths to demonstrate that the observations are indeed representative of the monthly-mean aerosols along the sampling tracks. There is no perfect way to perfume such a comparison. But for a minor comment on the screening of the data, I am satisfied with the approach.

The greater challenge for this paper is arriving at some generalized results that can guide the modelers. At the root of the challenge is that models may have many deficiencies that contribute to errors in the representation of the aerosol plumes, from errors in emissions to errors in transport, and uncertainties in the appropriate aerosol particle sizes and optical properties. There are only a few clues as to which errors might be contributing to the biases documented in the paper, so the end result is an illustration that all of these sources of model error contribute to causing a wide spread in the resulting aerosol distributions and physical properties among the models. This information is certainly worth sharing with the community, and this is exactly the kind of effort we should hope to see when we have high-quality datasets such as that from ORACLES. I think this paper would be suitable for publication if the authors can draw a stronger connection between the general limitations of the models discussed in the introduction as motivation for the paper and the results that they found. Thus, the discussion at the end of the paper should state how the results relate to specific shortcomings in the models in the literature as summarized in the manuscript. In the absence of drawing this connection, the paper just seems like a list of various model-data differences with no coherent interpretation or generalized outcome that the reader can take away from the study.

We appreciate the reviewer's comment. The models produce an almost surprising amount of diversity within their biases, and it is beyond the scope of this manuscript to attribute specific model shortcomings to the responsible processes. What does seem clear is that all of the models struggle with a realistic representation of the organic aerosol, which in turn may help explain the wide range in single-scattering-albedos and a mass extinction efficiency proxy between the models. We have edited the manuscript throughout to emphasize this. For example, the abstract now reads as:

"In the southeast Atlantic well-defined smoke plumes from Africa advect over marine boundary layer cloud decks; both are most extensive around September, when most of the smoke resides in the free troposphere. A framework is put forth for evaluating the performance of a range of global and regional atmospheric composition models against observations made during the NASA ORACLES (ObseRvations of Aerosols above CLouds and their intEractionS) airborne mission in September 2016. A strength of the comparison is a focus on the spatial distribution of a wider range of aerosol composition and optical properties than has been done previously. The sparse airborne observations are aggregated into approximately 2o grid boxes and into three vertical layers: 3-6 km, the layer from cloud top to 3 km, and the cloud-topped marine boundary layer. Simulated aerosol extensive properties suggest that the flight-day observations are reasonably representative of the regional monthly average, with systematic deviations of 30 % or less. Evaluation against observations indicates that all models have strengths and weaknesses, and there is no single model that is superior to all the others in all metrics evaluated. Whereas all six models typically place the top of the smoke layer within 0-500 m of the airborne lidar observations, the models tend to place the smoke layer bottom 300-1400 m lower than the observations. A spatial pattern emerges, in which most models underestimate the mean of most smoke quantities (black carbon, extinction, carbon monoxide) on the diagonal corridor between (6o E, 16o S) and (0o E, 10o S) in the 3-6 km layer, and overestimate them further south, closer to the coast, where less aerosol is present. Model representations of the above-cloud aerosol optical depth differ more widely. Most models overestimate the organic aerosol mass concentrations relative to those of

black carbon, and with less skill, indicating model uncertainties in secondary organic aerosol processes. Regional-mean free-tropospheric model ambient single scattering albedos vary widely, between 0.83-0.93 compared with in situ dry measurements centered at 0.86, despite minimal impact of humidification on particulate scattering. Modeled ratio of the particulate extinction to the sum of the black carbon and organic aerosol mass concentrations (a mass extinction efficiency proxy) are typically too low and vary too little spatially, with significant inter-model differences. Most models overestimate the carbonaceous mass within the offshore boundary layer. Overall, the diversity in the model biases suggests that different model processes are responsible. The wide range of model optical properties requires further scrutiny because of their importance for radiative effect estimates."

Overall our study is limited to a documentation of model biases, with error attribution, and left to future studies. Although we highlight a few errors common to all of the models, the model diversity suggests that the underlying shortcomings may differ between the models.

Other comments:

The abstract claims a "new approach to utilizing airborne aerosol measurements", but is not explicit about what aspect of the study the authors are claiming is new.

The abstract, provided above, has been modified to detail the approach.

*Is there a citation or other evidence to support the use of "altitudes below (RH(%)- 60)*40m to define the boundary layer depth?*

No, there isn't. While the vertical gradient in temperature or water vapor mixing ratio would determine the boundary layer more accurately, airborne data only occasionally provide it. The

formula was empirically derived from the collective RH profiles shown in Figure 2. This indicates the close correspondence of RH to boundary layer depth for this time period.

The grey points in figure 2 are apparently observational values that could not be successfully placed in one of the three altitude classification. I presume these data are not included in the comparison with the model. Is there a sampling bias related to this? In particular I would think that the low altitude data points shown in grey, presumably corresponding to cases where the top of the boundary layer is too difficult to discriminate, do represent a condition that happens with some regularity. Shouldn't the models reproduce a similar condition occasionally?

As the reviewer points out, most of the grey data points refer to the inversions observed at the top of the boundary layer and are excluded. We neglect them as they represent less than 3 % of the observations, compared to 48 % in 3-6 km, 21 % in FT<3 km and 17 % in the MBL, 11 % above 6 km.

Each of the inversions is less than 100 m deep. The model products also exclude inversions from both the boundary layer and the free troposphere. As their vertical resolution is not fine enough to represent the gradient over such narrow depths, inversions are represented as a step function with zero depth.

Can the authors draw some connections between the systematic biases in the thickness of the aerosol layer and the extinction optical properties of the particles? Are there some known deficiencies in the aerosol radiative forcing or fluxes of any of these models that could be tied to the biases in plume thickness and optical properties reported by the authors? Do the biases the authors have found tend to reinforce one another in magnifying errors in the bulk radiative effect of aerosols, or perhaps are there some compensating errors? Answering these questions would help clarify what has been learned from quantifying all of these biases.

Our results provide no systematic evidence that an overestimate of the aerosol layer geometrical thickness is accompanied by an exaggerated vertical dilution of aerosols, as witnessed by the model diversity in ACAOD. Our results, however, leave the possibility that compensation, or magnification, could happen on a model-by-model basis. We do include a new Section 7.3, reproduced within the response to Reviewer 1, that discusses how the model biases documented within this study could impact the model aerosol radiative effect estimate.

The clearest result we have found is that the models have difficulty in representing the fractional composition of the aerosol, with generally too much organic aerosol for the same amount of BC. This has ramifications for all of the optical properties. While we cannot prescribe model remedies, the comparison overall does suggest that more focus on the model representation of the organic aerosol processes may also lead to improvements in the model optical properties. The wide range of model biases, however, preclude us from making broader statements than this. An upcoming companion paper by Doherty et al., will help draw the connection between the aerosol biases and their impact on the radiation fields, which are also dependent upon the representation of the underlying cloud field.